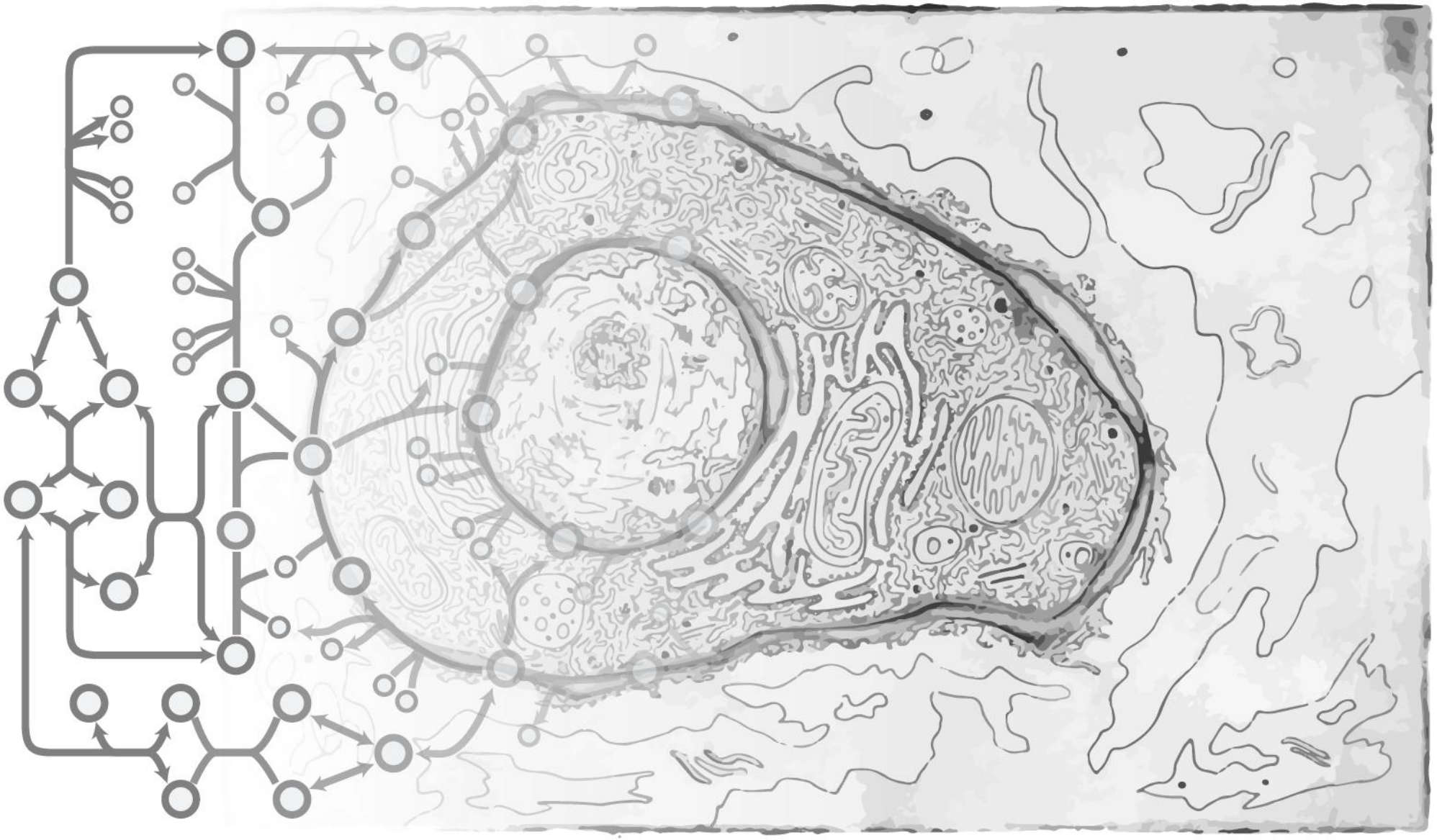
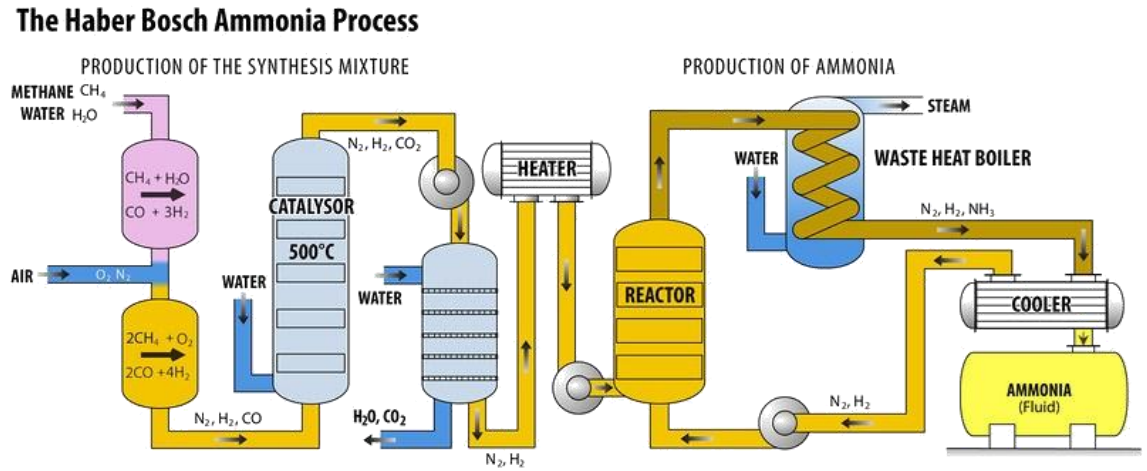


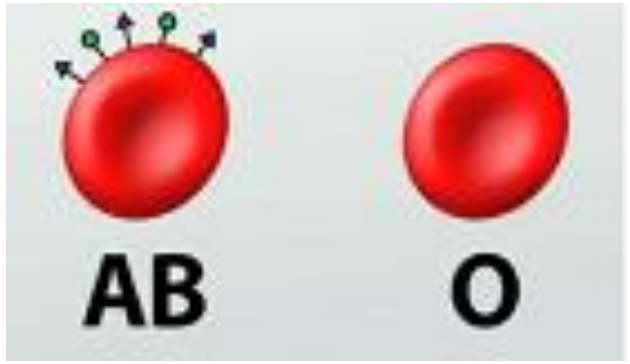
ENGINEERING CELL FACTORIES WITH MAMMALIAN SYSTEM BIOLOGY, OMICS, BIG DATA

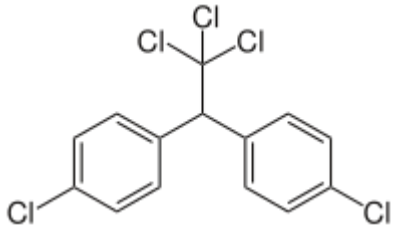
Nathan E. Lewis, PhD
University of California, San Diego
March 8, 2024

INVENTIONS THAT HAVE SAVED THE MOST LIVES

1. The Haber Bosch process
2. Blood groups and blood transfusion technology
3. Green revolution
4. Clean water
5. Vaccines
6. Antibiotics
7. Oral rehydration therapy
8. DDT
9. Insulin



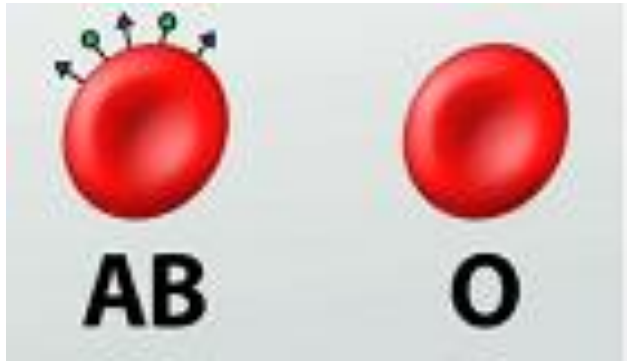




INVENTIONS THAT HAVE SAVED THE MOST LIVES

1. The Haber Bosch process
2. Blood groups and blood transfusion technology
3. Green revolution
4. Clean water
5. Vaccines
6. Antibiotics
7. Oral rehydration therapy
8. DDT
9. Insulin

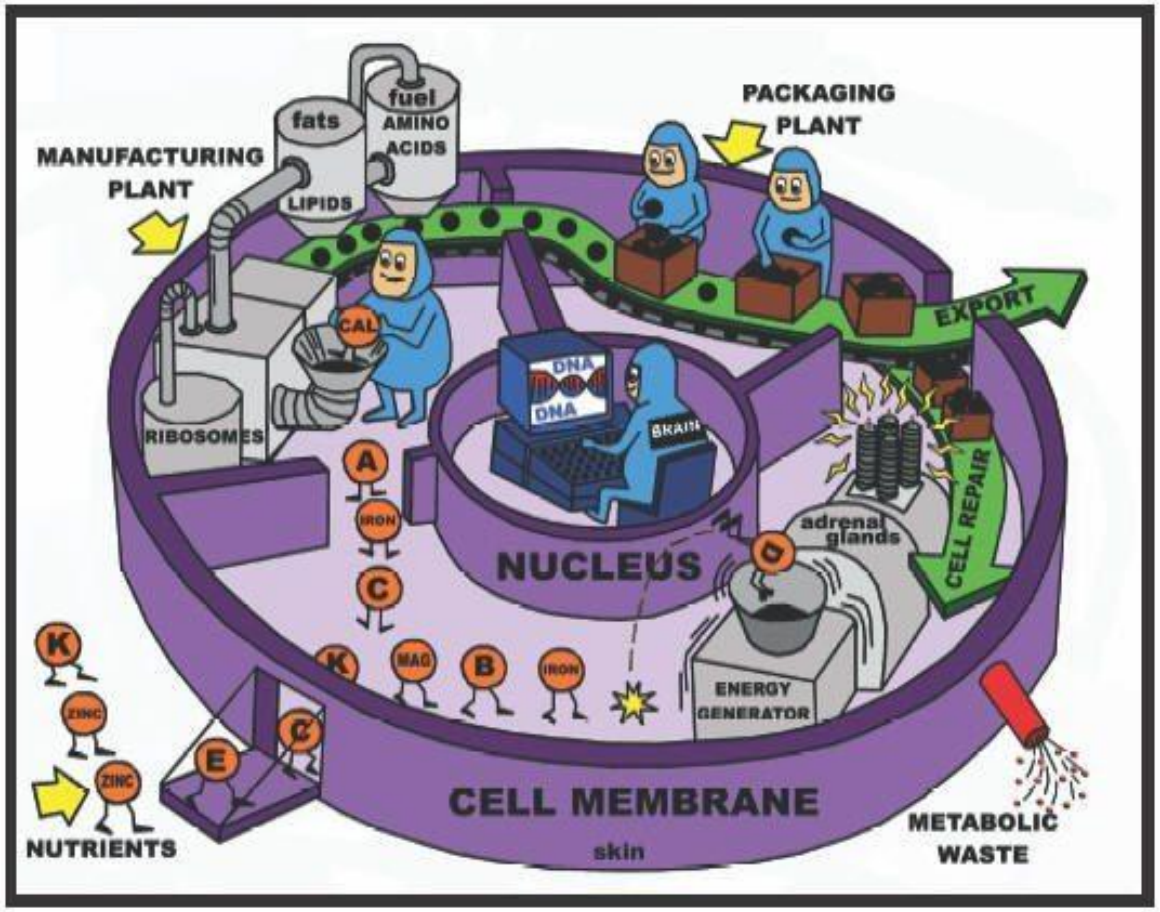
(Biotech)



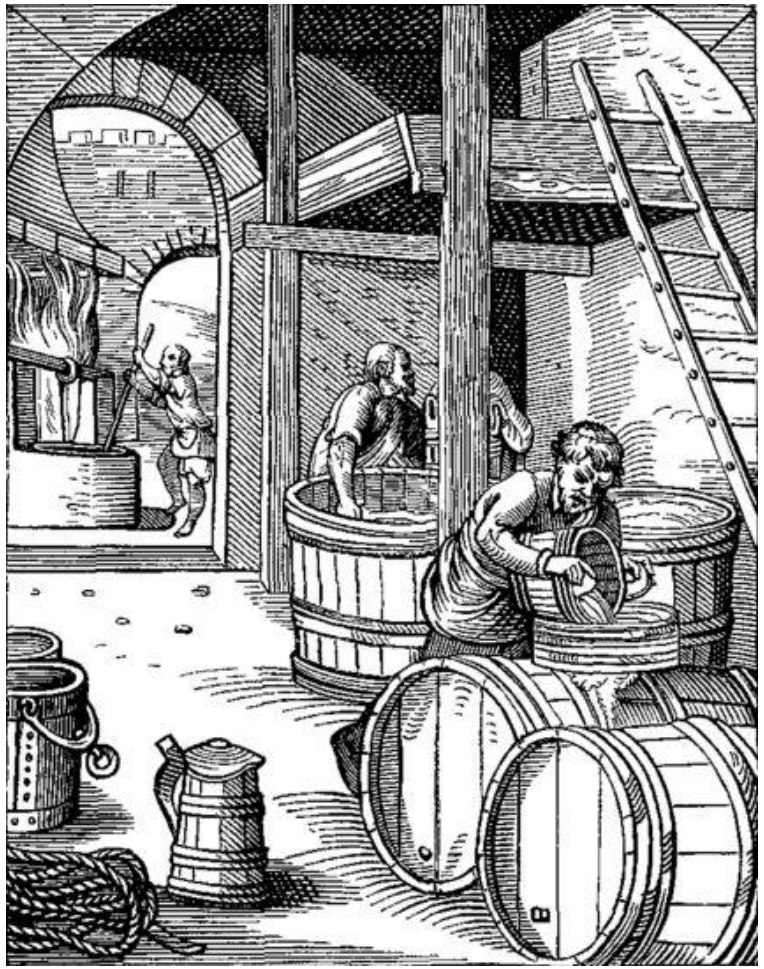
INVENTIONS THAT HAVE SAVED THE MOST LIVES

1. The Haber Bosch process
2. Blood groups and blood transfusion technology
3. Green revolution
4. Clean water
5. Vaccines
6. Antibiotics
7. Oral rehydration therapy
8. DDT
9. Insulin

(Biotech)



BIOTECHNOLOGY THROUGH CENTURIES



Biotechnology – using a biological system to make products. (16th century)

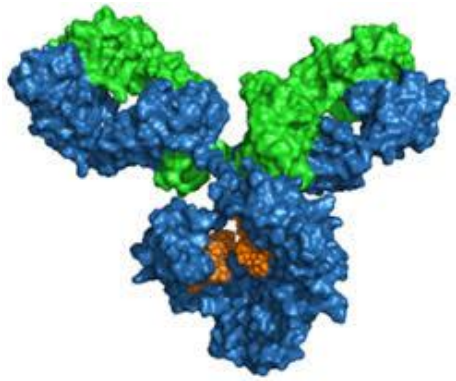
Technology

Bioreactors for producing proteins, NRC Biotechnology Research Institute, Montréal, Canada

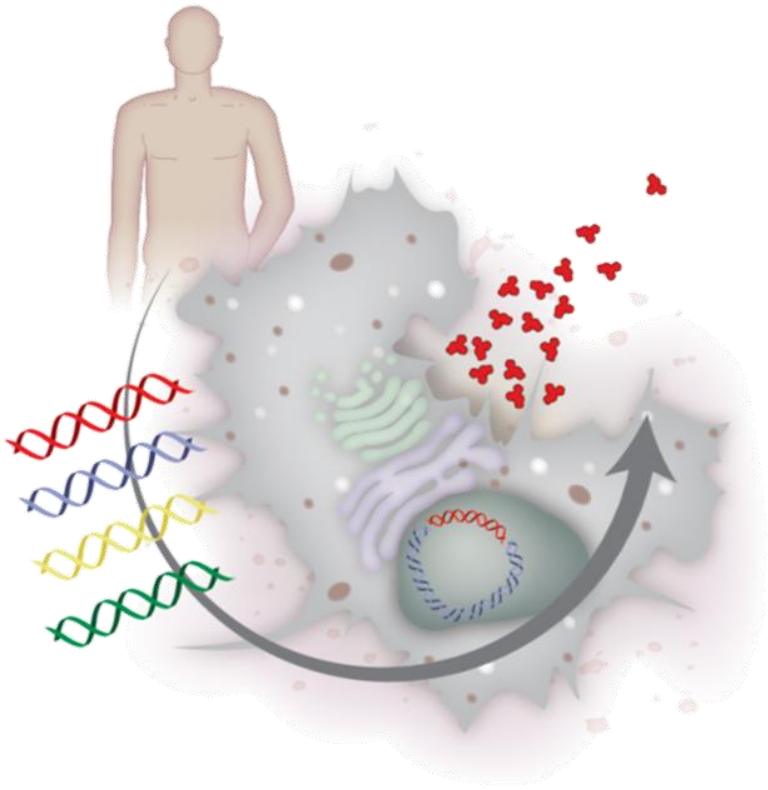
Cell Engineering

Biotechnology – using an **engineered** biological system to make products. (21th century)

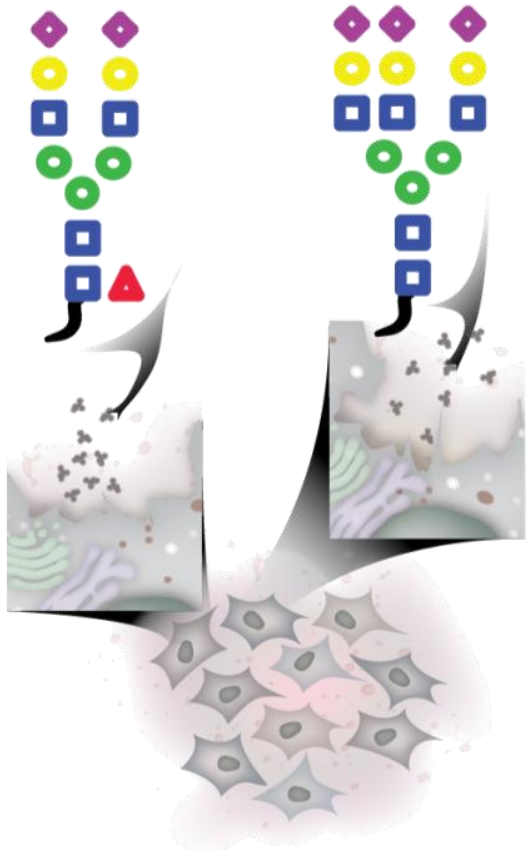
UNIQUE CHALLENGES FOR MAMMALIAN BIOPRODUCTION



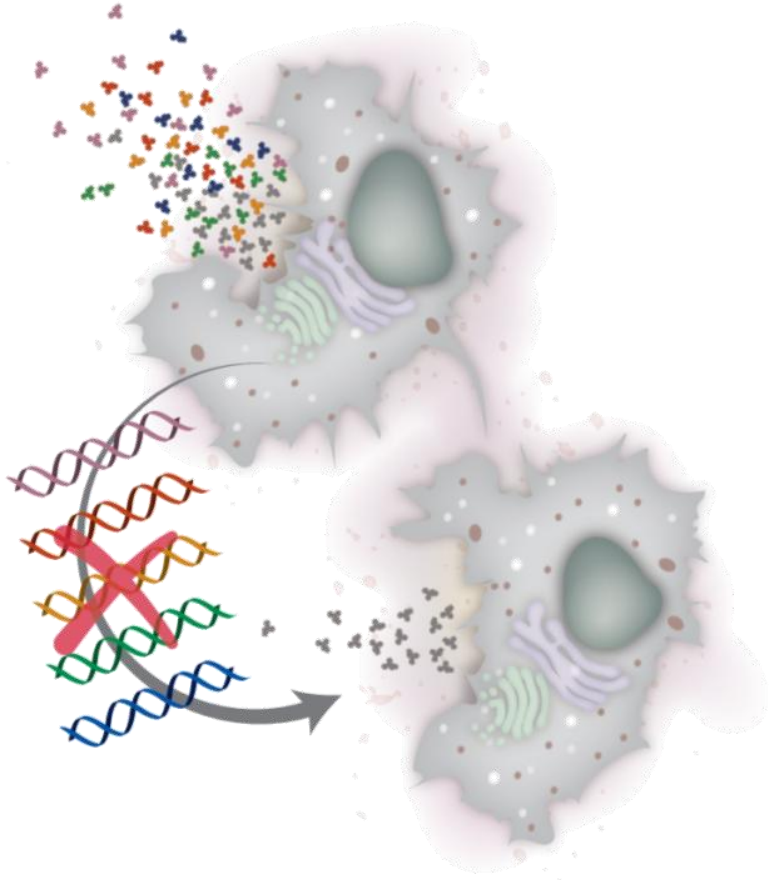
High yield



Controlled Quality Attributes



Purity



OVERVIEW

Context

Observational strategies

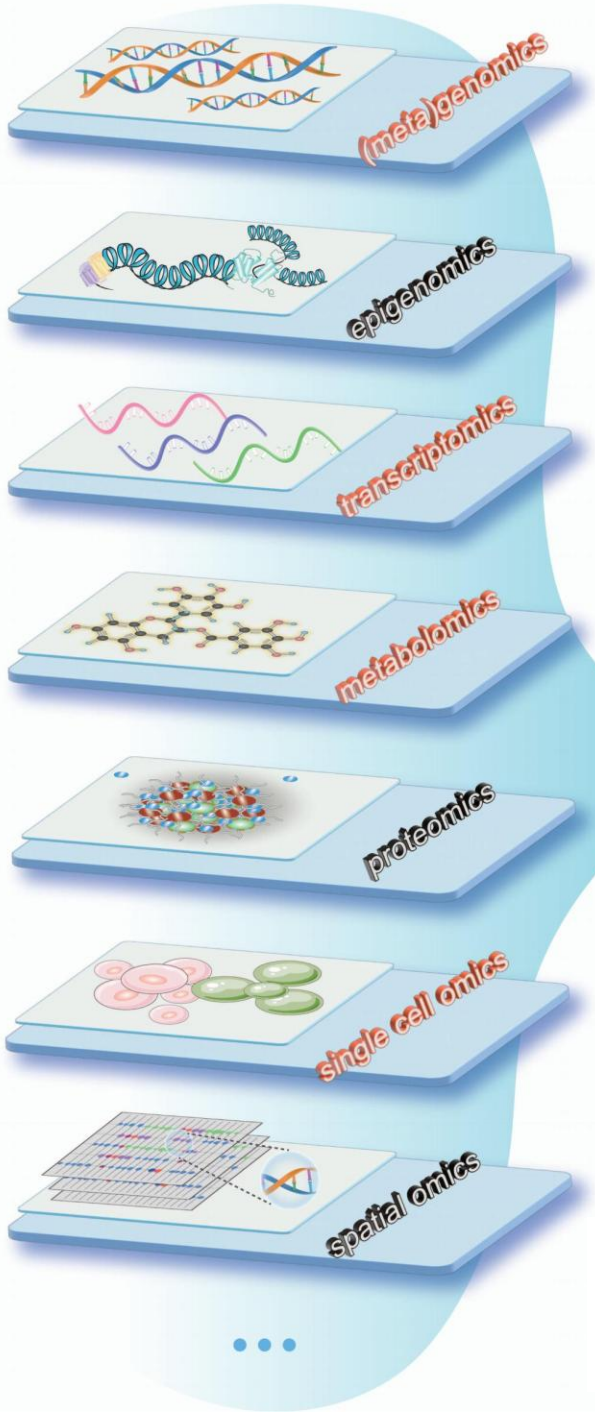
How to measure parts

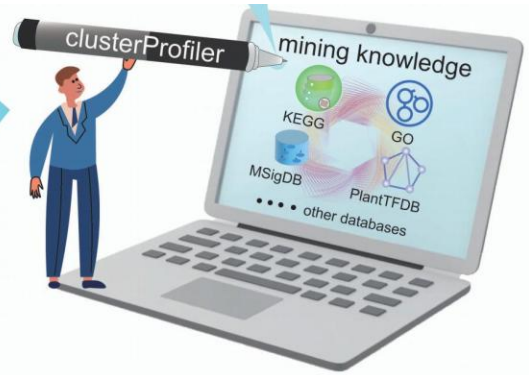
- Nucleic acids (DNA/RNA)
- Protein
- Metabolites

How to measure interactions

- Protein-protein
- Protein-DNA
- Post-translational modifications

Interventional strategies



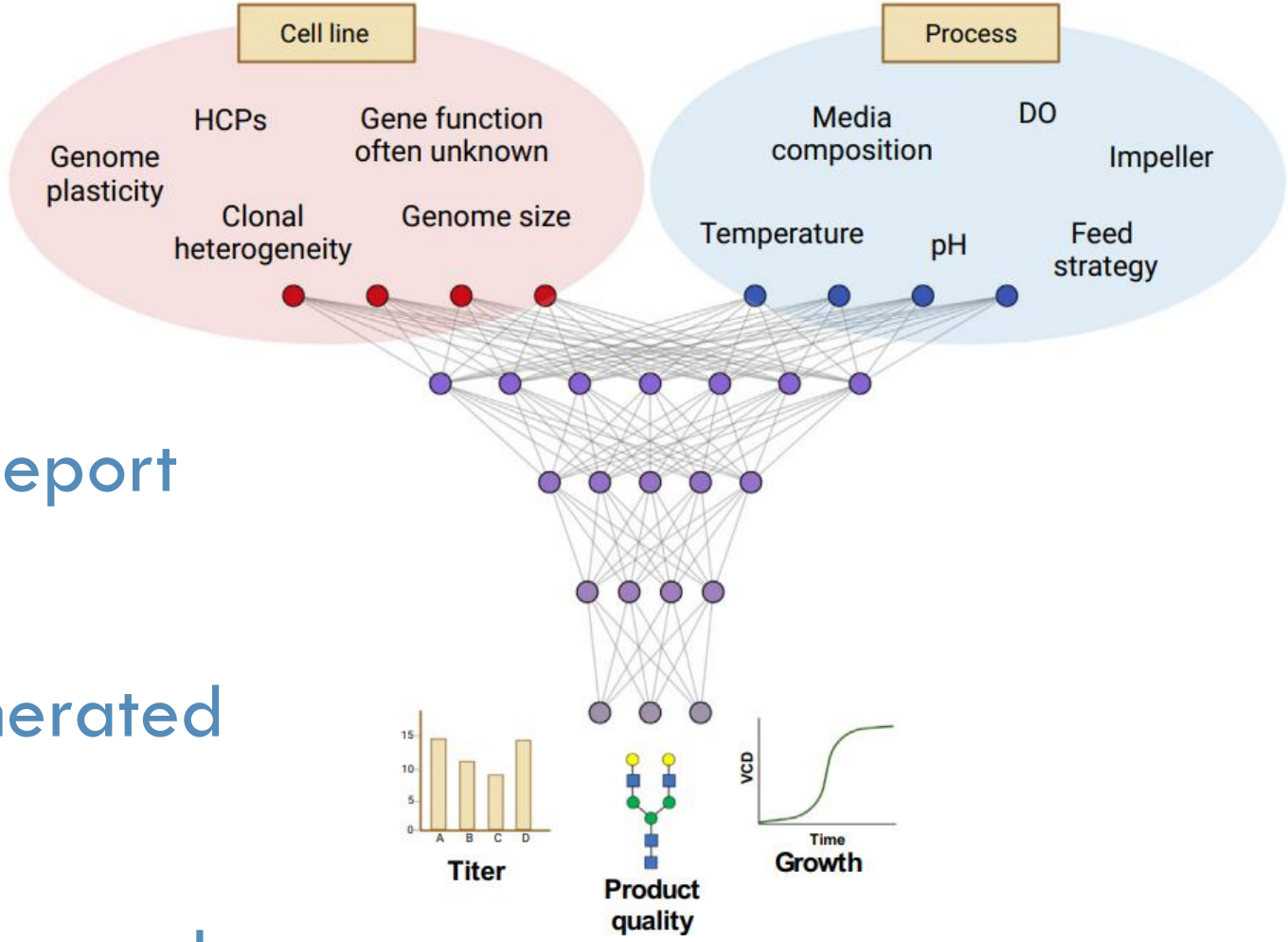


Using clusterProfiler to characterize multiomics data

Shuangbin Xu, Erqiang Hu, Yantong Cai, Zijing Xie, Xiao Luo, Li Zhan, Wenli Tang, Qianwen Wang, Bingdong Liu, Rui Wang, Wenqin Xie, Tianzhi Wu, Liwei Xie & Guangchuang Yu

Nature Protocols **19**, 3292–3320 (2024) | Cite this article

WHERE DOES TODAY'S LECTURE FIT?



Not necessary for the final report

Understand how data is generated

Understand how data can be used

Review > Trends Biotechnol. 2023 Sep;41(9):1127-1138. doi: 10.1016/j.tibtech.2023.03.009. Epub 2023 Apr 14.

From observational to actionable: rethinking omics in biologics production

Helen O Masson [1], Karen Julie la Cour Karottki [2], Jasmine Tat [3], Hooman Hefzi [4], Nathan E Lewis [5]

HOW DO CELLS WORK? LESSONS FROM A RADIO

The Regency TR-1 was first portable and practical transistor radio

Phenotype:

- On or off
- Played a spectrum of different AM stations

GENOTYPE: A CATALOG OF CELL PARTS AND THEIR FEATURES

The Regency TR-1 Genotype:

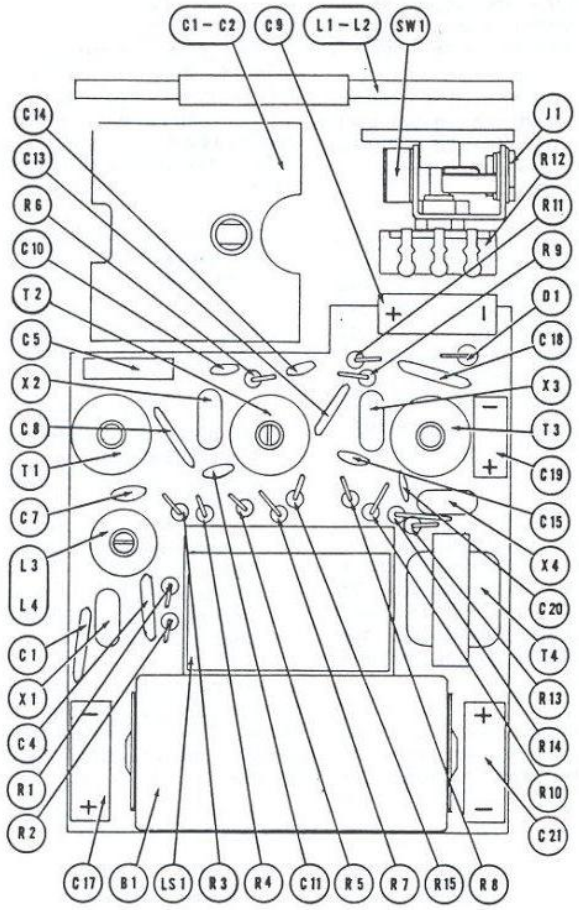


Fig. 6. Top Chassis Components in Regency Model TR-1.

PARTS LIST.

CAPACITORS

Symbol No.	Part No.	Description	Type
C1, C18	20-075-23	.02 Mfd 1/4" Dia. -20 + 80%	Disc Ceramic
C4	20-075-21	.01 Mfd 9/16" Dia. -20 + 80%	Disc Ceramic
C5	100-771	268 Mmfd 2%	Silver Mica
C7, C11, C15, C20	20-075-21	.001 Mfd 1/4" Dia. G. M. V.	Disc Ceramic
C8, C13	100-772-1	.05 Mfd 9/16" Square -20 + 80%	Disc Ceramic
C9, C22	300-473-2	40 Mfd / 3 Volt	Electrolytic
C17	300-473-3	5 Mfd / 25 Volt	Electrolytic
C19	300-473-1	2 Mfd / 3 Volt	Electrolytic

RESISTORS

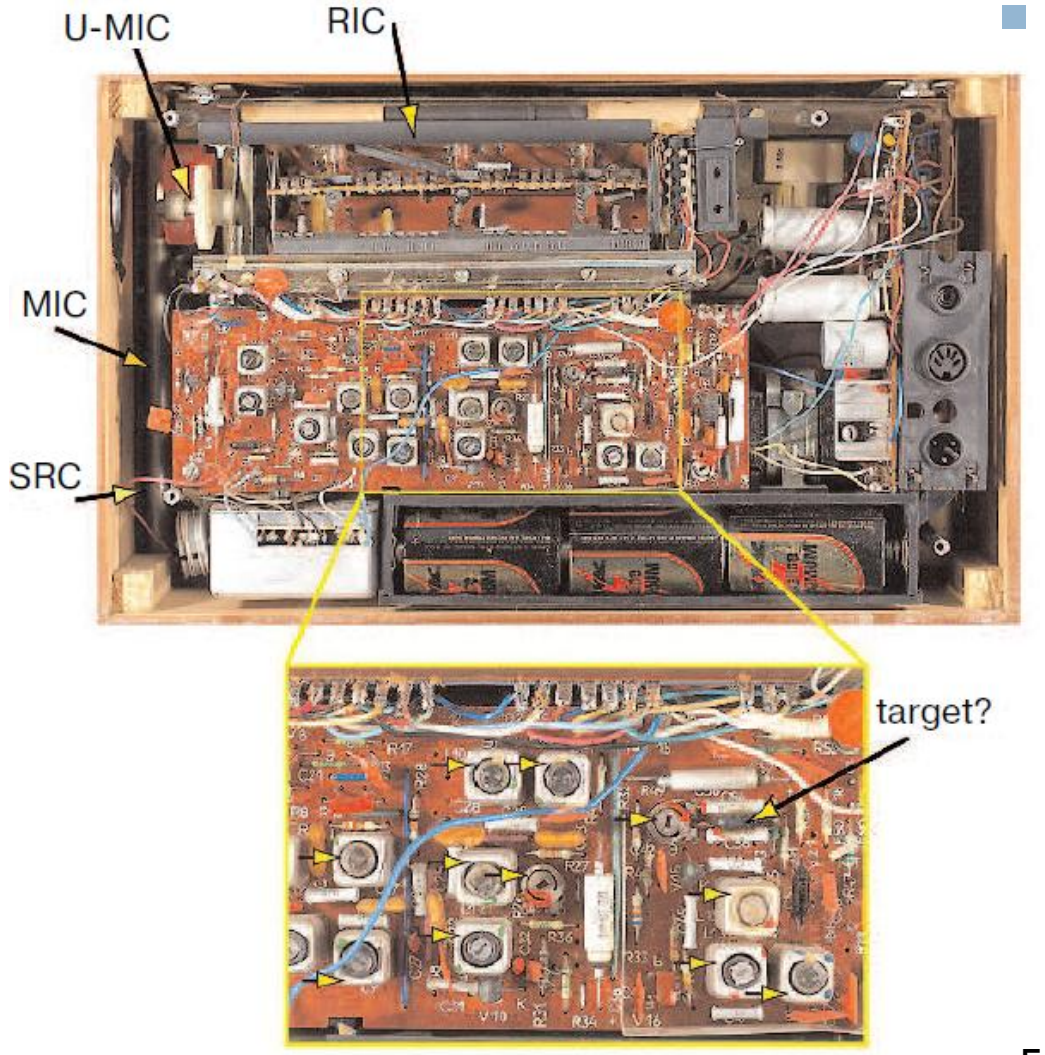
Symbol No.	Description
R1	470 K 1/4 Watt - 20% Carbon
R2	10 K 1/4 Watt - 20% Carbon
R3, R7, R10	2.2 K 1/4 Watt - 20% Carbon
R4	100 K 1/4 Watt - 10% Carbon
R5, R6, R9	560 Ohm 1/4 Watt - 10% Carbon
R8, R11	2.7 K 1/4 Watt - 10% Carbon
R13	3.9 K 1/4 Watt - 10% Carbon
R14	33 K 1/4 Watt - 10% Carbon
R15	1 K 1/4 Watt - 10% Carbon

MISCELLANEOUS

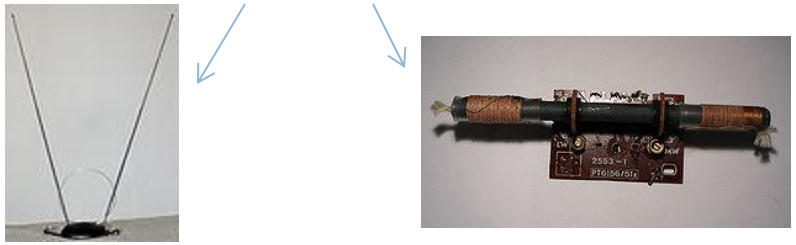
Symbol No.	Part No.	Description
B1		Regency No. 215 Battery
C1, C2	300-433	Variable Tuning Capacitor
D1	100-773	Germanium Diode Detector
J1	300-434	Earphone Jack
L1, L2	300-432	Ferrite Core Loop Antenna
L3, L4	100-731	Oscillator Coil
LS1	100-766	Loud Speaker
R12	100-596-2	Volume Control (1K-Audio Taper)
T1, T2, T3	100-728	IF Transformer
T4	100-629	Output Transformer
X1	100-626	Mixer Transistor
X2, C10	100-767	IF Transistor Replacement (Transistor & Neut. Cap.)
X3, C11		
X4	100-628	Audio Transistor
	600-061	Case (Front & Back-State Color)
	100-732	Dial
	100-734	Dial Screw
	300-435	Volume Control Knob (State Color)

- Determining the parts (omics):
  - Open one up and classify the parts by color, size, shape, material

BIOCHEMISTRY AND GENETICS: DETERMINING PART FUNCTIONS



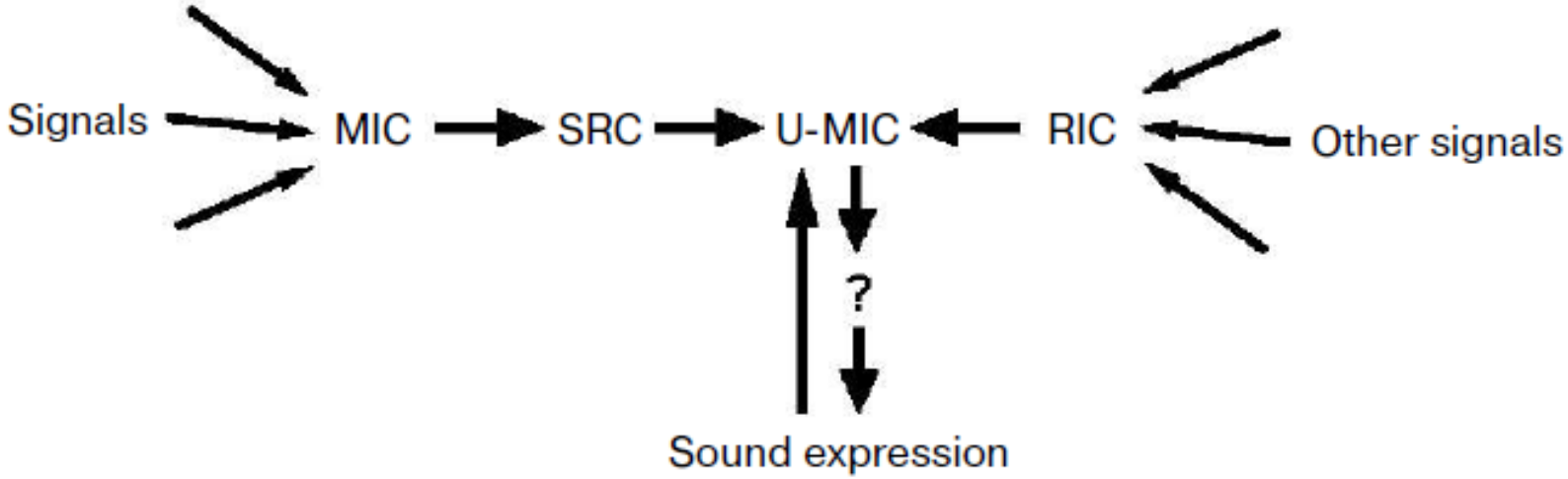
- Determining the function of parts
  - Buy many (with inflation = ~$360 each) and shoot them for phenotypes (KO screen)
  - Test properties of each part (biochemical assays)
  - Name each component (SRC, MIC, RIC, and U-MIC)



From Lazebnik, "Can a biologist fix a radio?"

FINDING THE CELLULAR WIRING DIAGRAM LINKING GENOTYPE TO PHENOTYPE

A biologist linking the parts of a Regency TR-1

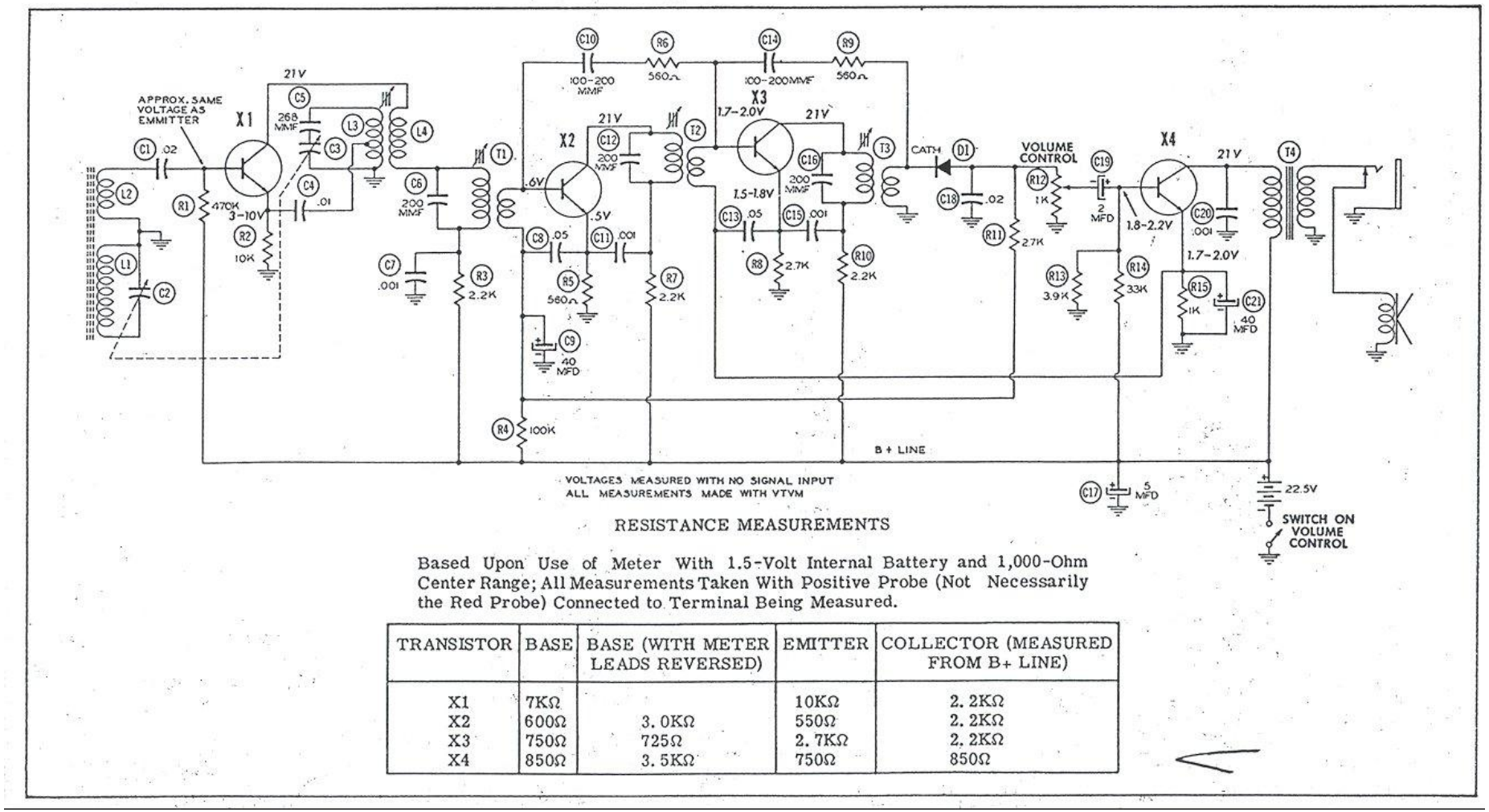


- To refine the wiring diagram, smash even more radios, then pull on each piece to find all components that sticks to it (ChIP, co-IP)

From Lazebnik, "Can a biologist fix a radio?"

FINDING THE CELLULAR WIRING DIAGRAM LINKING GENOTYPE TO PHENOTYPE

An engineer linking the parts of a Regency TR-1



RESISTANCE MEASUREMENTS

Based Upon Use of Meter With 1.5-Volt Internal Battery and 1,000-Ohm Center Range; All Measurements Taken With Positive Probe (Not Necessarily the Red Probe) Connected to Terminal Being Measured.

TRANSISTOR	BASE	BASE (WITH METER LEADS REVERSED)	EMITTER	COLLECTOR (MEASURED FROM B+ LINE)
X1	7KΩ		10KΩ	2.2KΩ
X2	600Ω	3.0KΩ	550Ω	2.2KΩ
X3	750Ω	725Ω	2.7KΩ	2.2KΩ
X4	850Ω	3.5KΩ	750Ω	850Ω

WHAT TOOLS ARE NEEDED TO CONNECT A CELL'S GENOTYPE AND PHENOTYPE?

Component list

Component functions

Component interactions

Knowledge on how interactions and states affect function

A standardized quantitative "language" for describing all interactions and simulating function

- Language = math, logical rules, wiring diagrams

REDUCTIONISM VS. SYSTEMS ANALYSIS

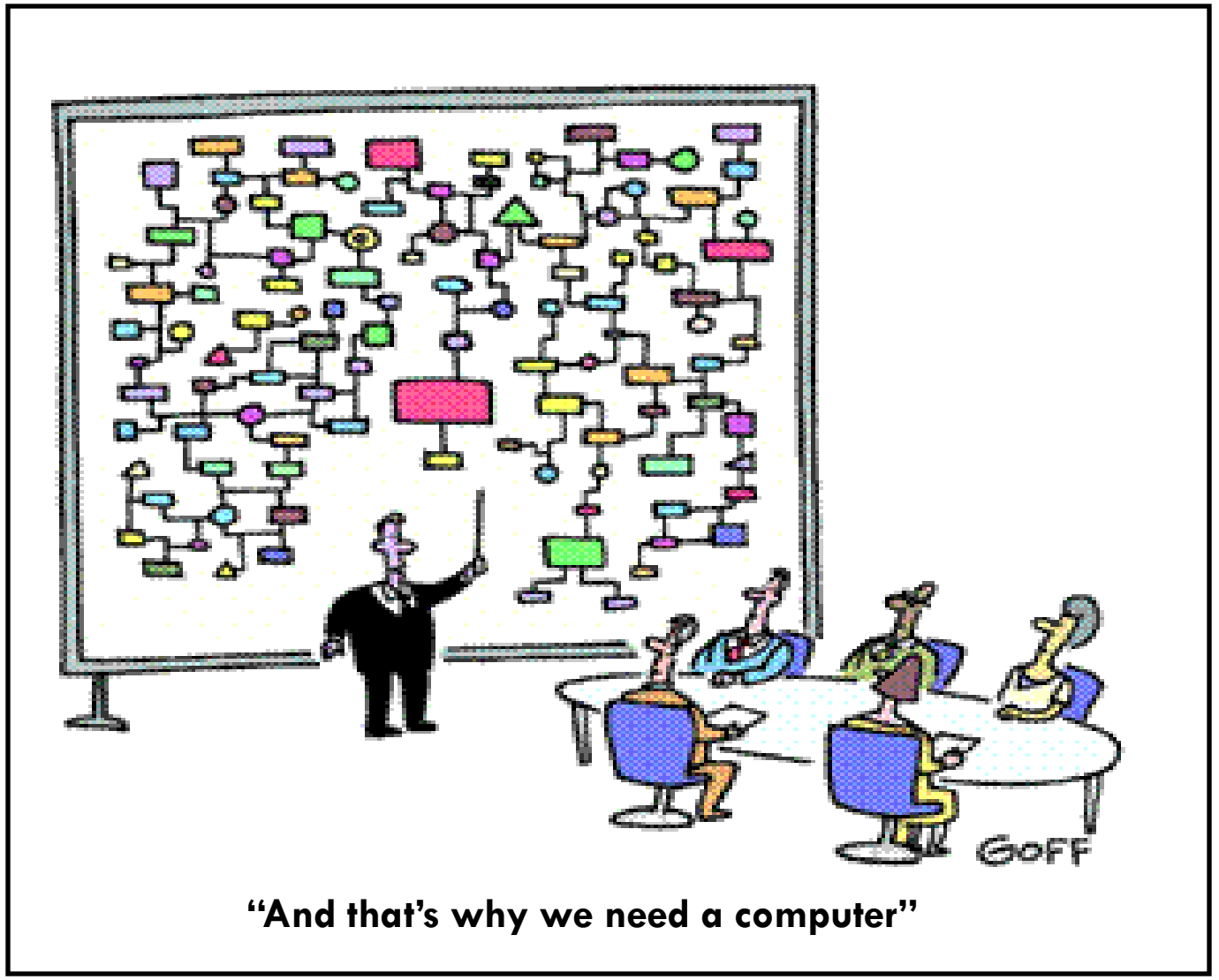
Reductionism

- You only need to know the behavior of components
- In complex systems, no new phenomena will emerge when we consider interactions in the entire system
- Clearly a flawed concept

Systems approach

- Elements (network nodes) and their interactions (network links)
- Sometimes details of each element become less important than the network interactions
- Important systems-level properties include robustness, fragility, modularity, hierarchy, evolvability, redundancy
- Most importantly, components function in a network context

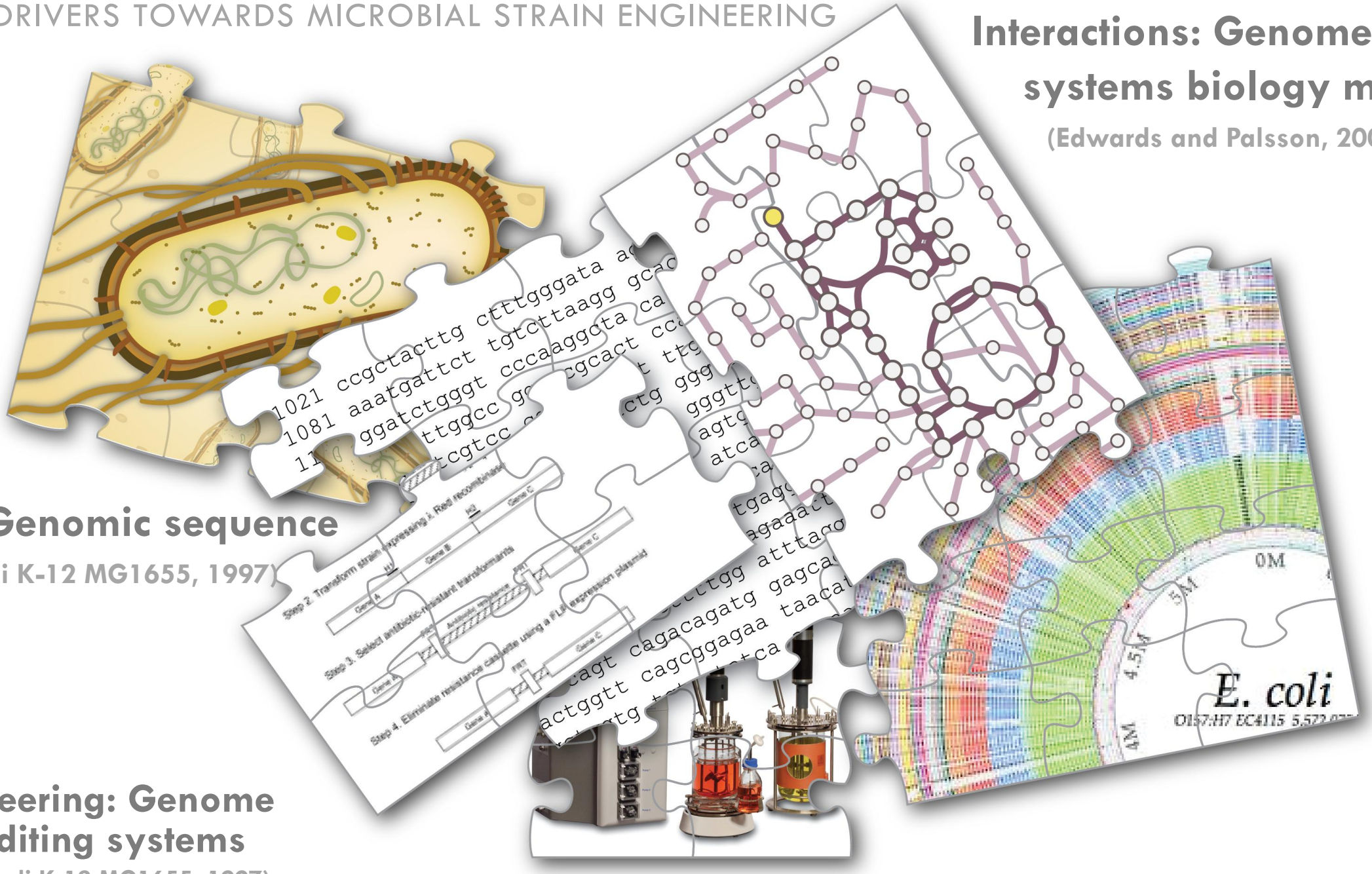
BUT… THERE ARE MANY COMPONENTS, AND EVEN MORE INTERACTIONS



THREE DRIVERS TOWARDS MICROBIAL STRAIN ENGINEERING

Interactions: Genome-scale systems biology models

(Edwards and Palsson, 2000)



Parts: Genomic sequence

(E. coli K-12 MG1655, 1997)

Engineering: Genome editing systems

(E. coli K-12 MG1655, 1997)

Three drivers towards a more complete picture in CHO cell engineering

Parts: Genomic sequences

(Xu, 2011; Lewis, 2013; Brinkrolf, 2013)

Proteome

(Baycin-Hizal, 2012)

Transcriptome

(Baycin-Hizal, 2012)

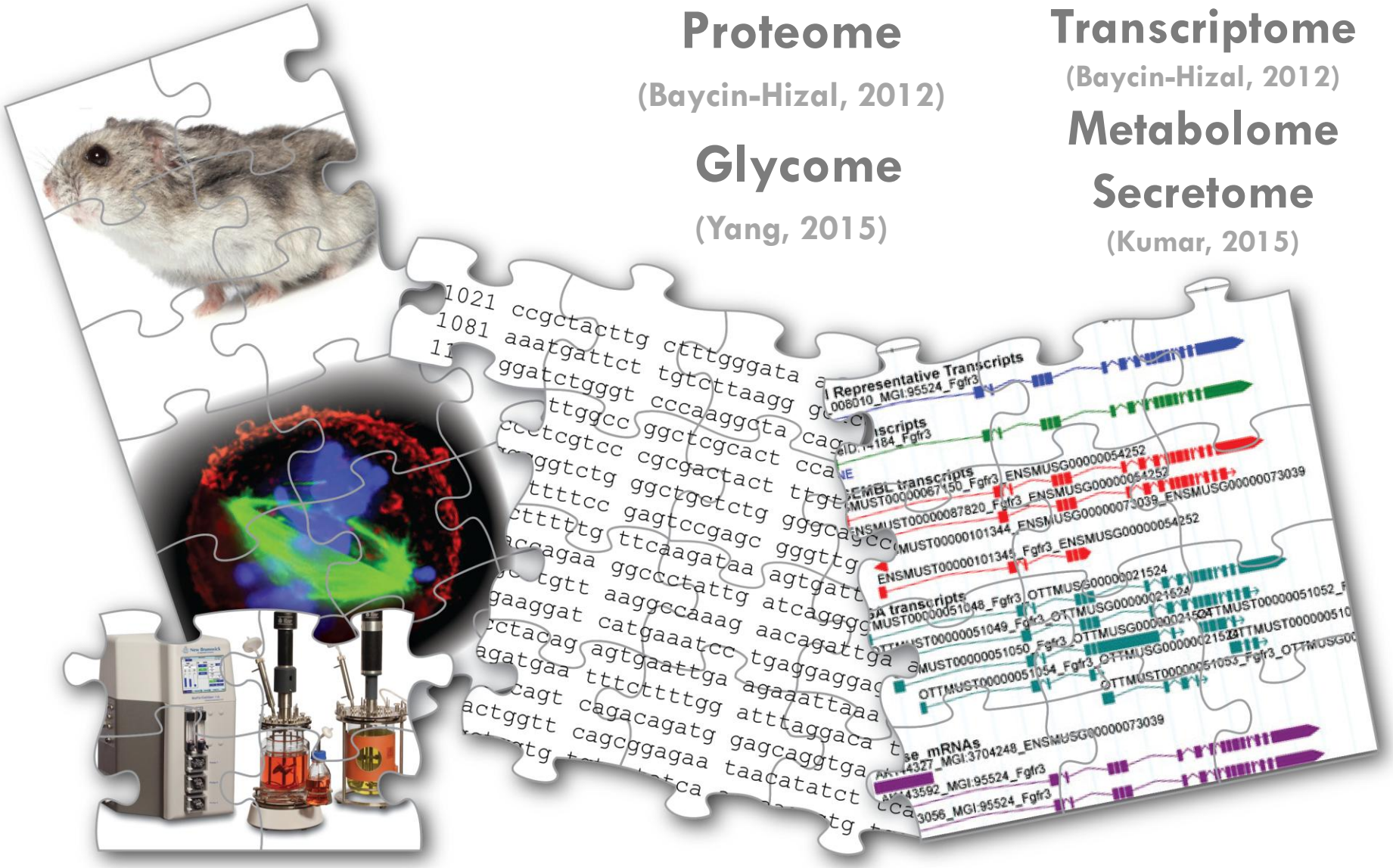
Metabolome

Glycome

(Yang, 2015)

Secretome

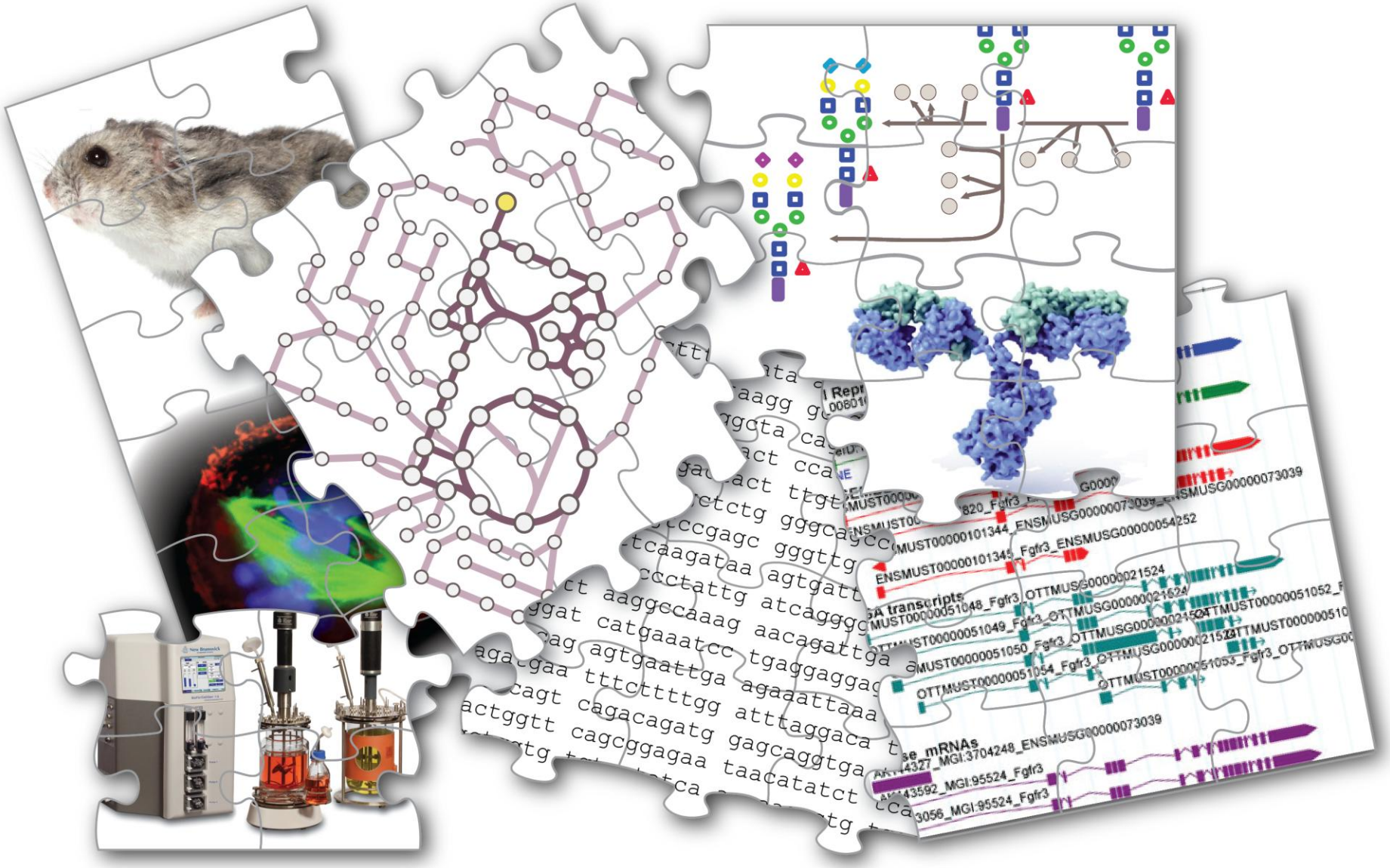
(Kumar, 2015)



Three drivers towards a more complete picture in CHO cell engineering

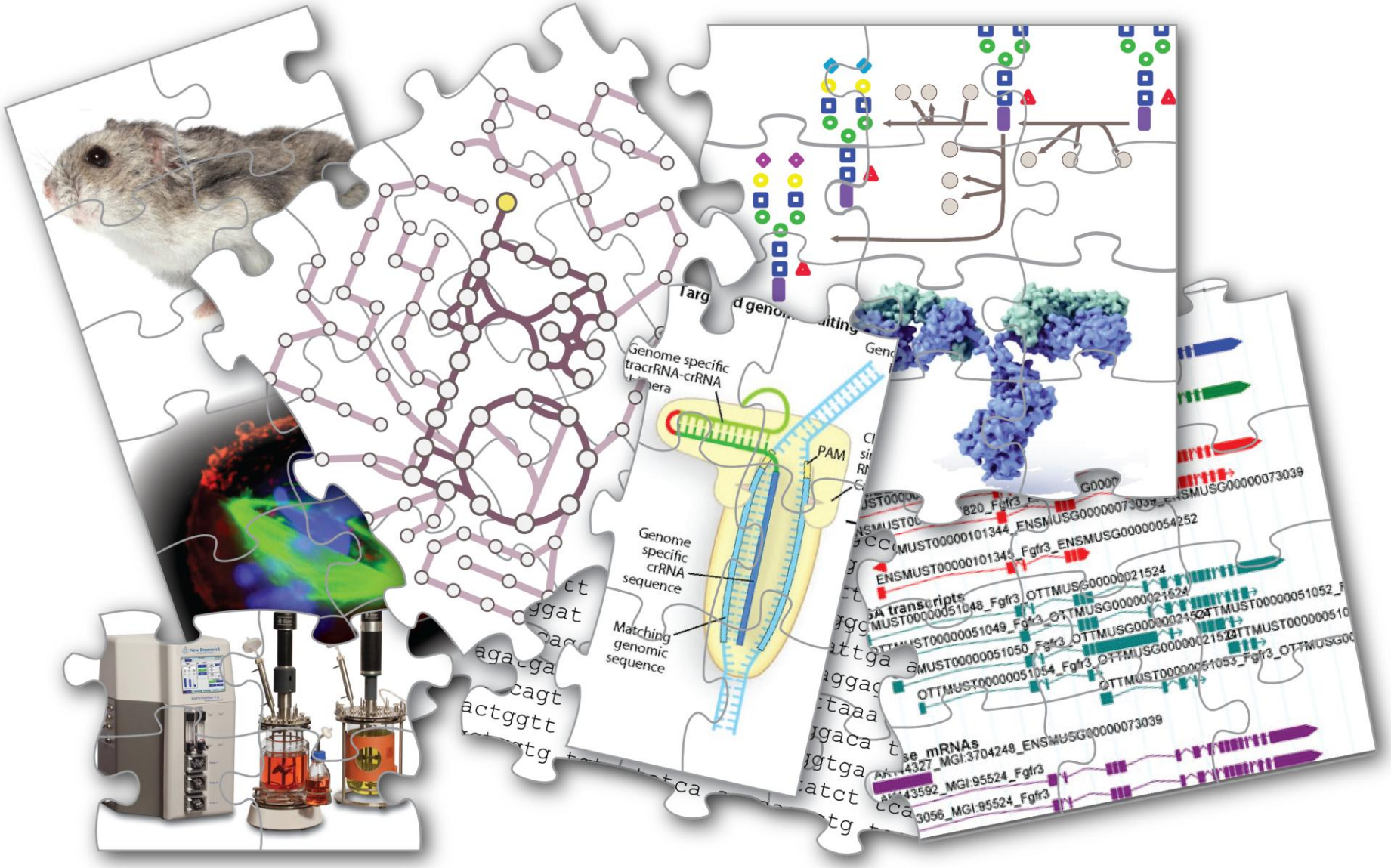
Interactions: Genome-scale Systems biology models

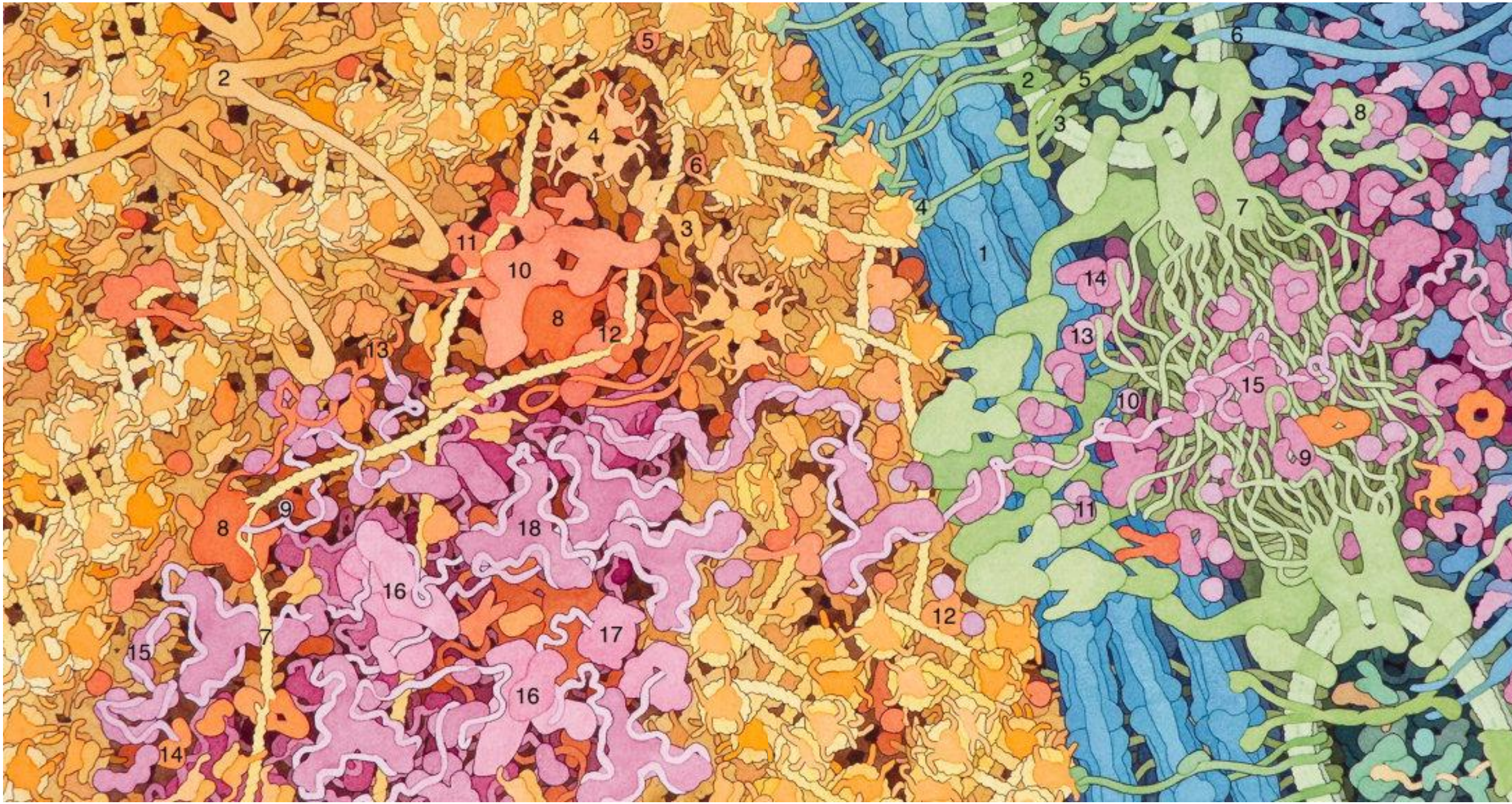
(Hefzi, 2016; Spahn, 2016; Gutierrez, submitted)



Three drivers towards a more complete picture in CHO cell engineering

Engineering: ZFNs, TALENs, CRISPR/CAS9

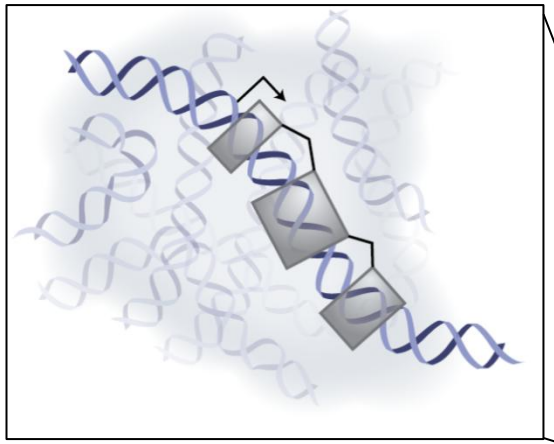




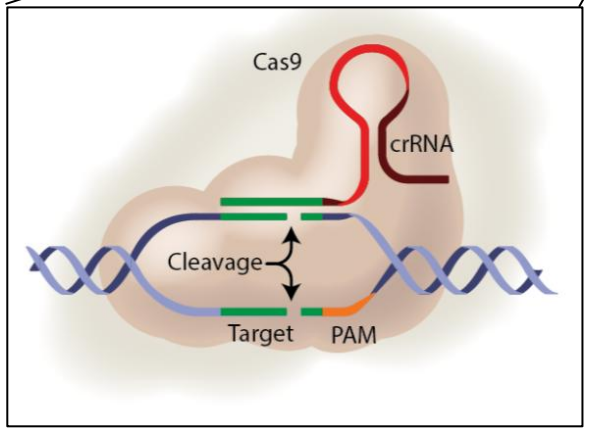
David Goodsell, TSRI

CONTROLLING PROTEIN PRODUCTION IN CHO CELLS

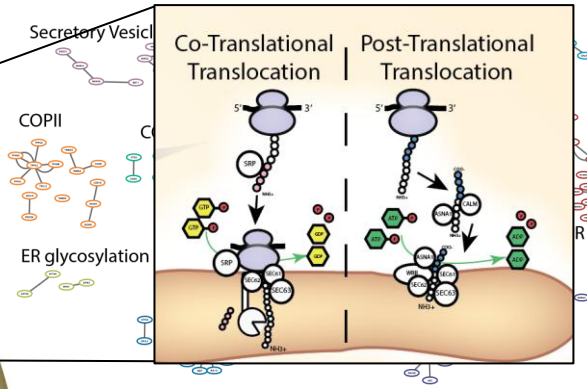
Parts: Genomic sequences

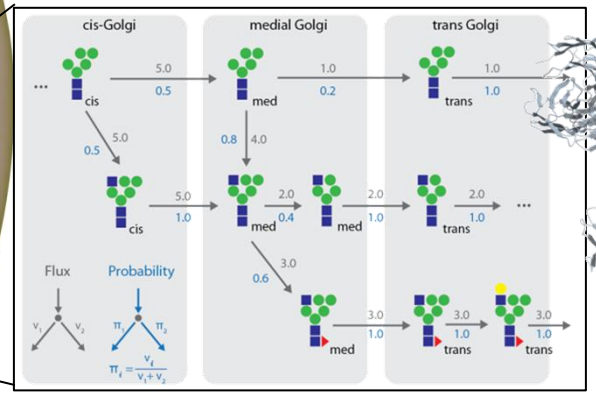


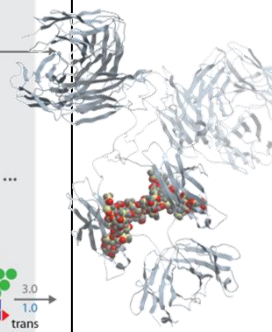
Engineering: CRISPR, DNA synthesis, etc.

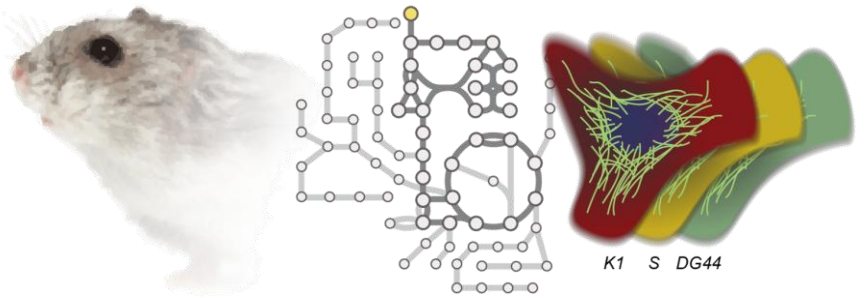


Interactions: Systems biology models









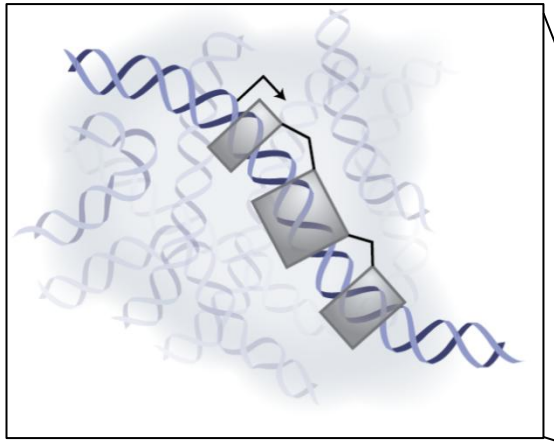
Amino acids

Sugar nucleotides

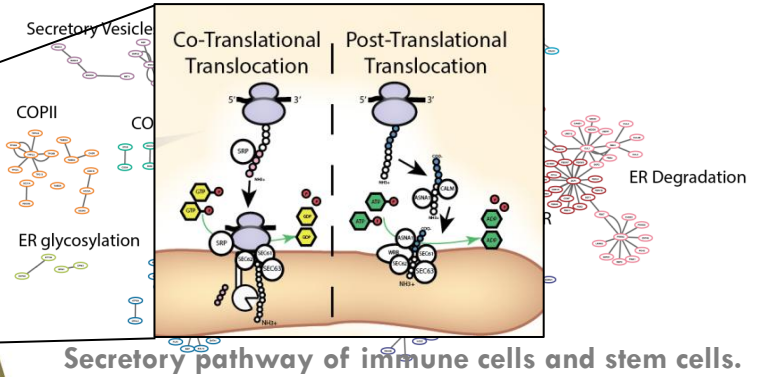
Parts: Genomic sequences

Interactions: Systems biology models

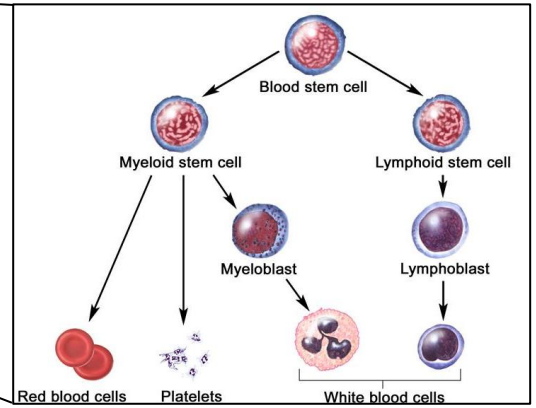
TOWARDS UNDERSTANDING AND ENGINEERING CELL THERAPY CELLS



Whole genome analysis of engineered cells
Sequence / assembly of antibody genes

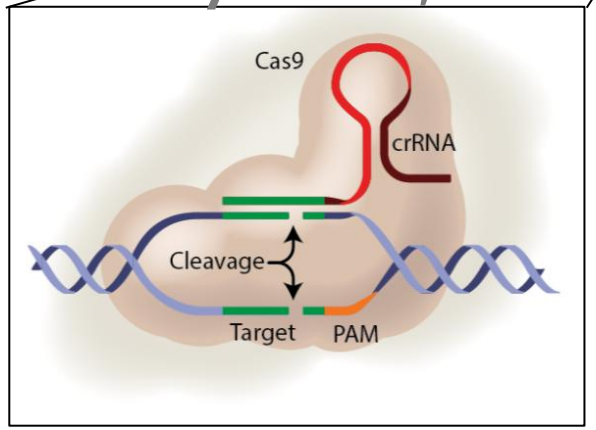


Secretory pathway of immune cells and stem cells.
Study of cell communication.

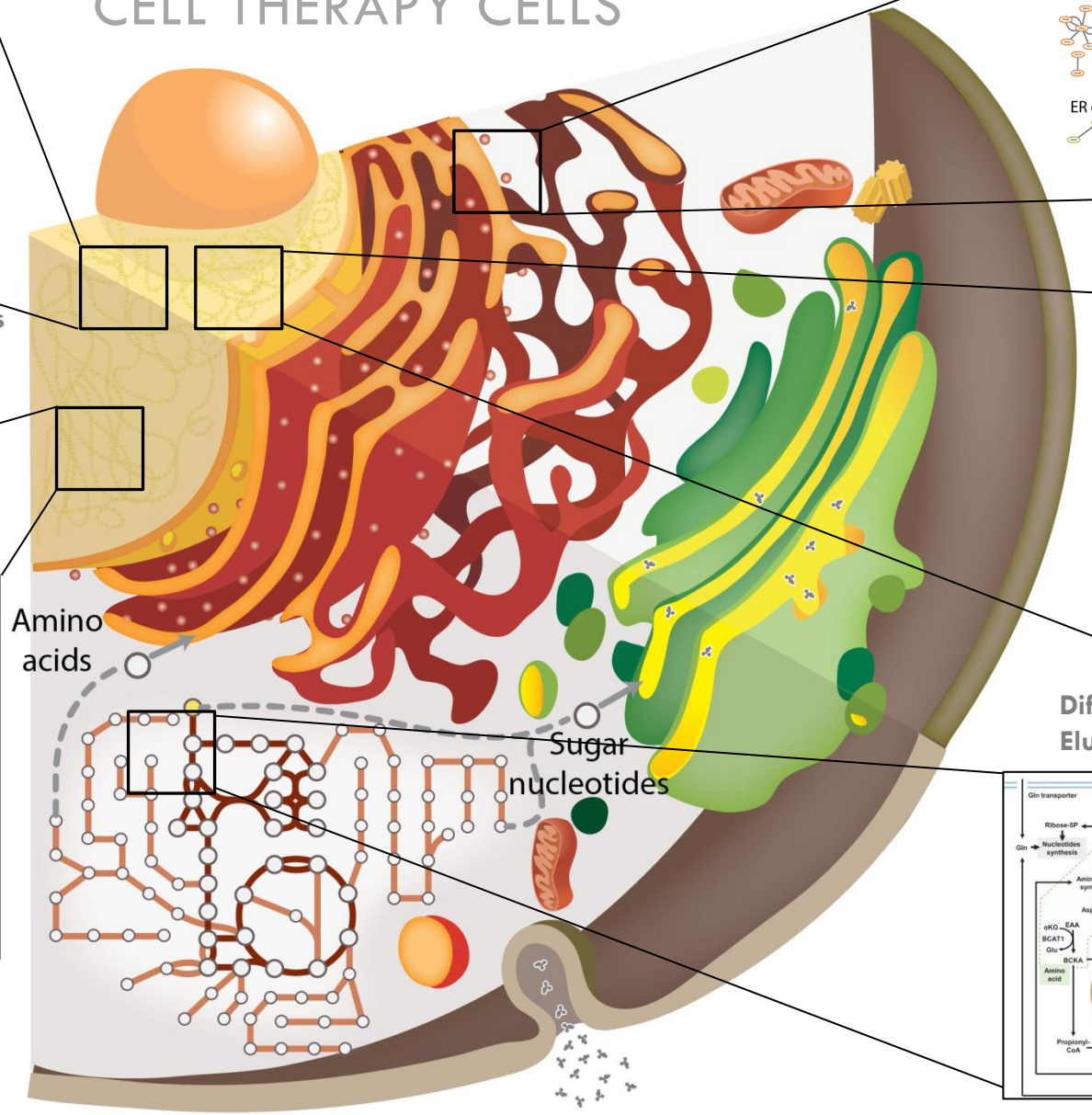


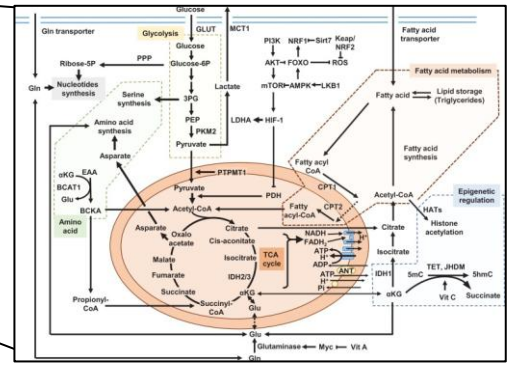
Differentiation pathway of hematopoietic cell lineages.
Elucidation of secreted developmental factors

Engineering: CRISPR, DNA synthesis, etc.



CRISPR screens
RNA editing
Regulatory switches





Stem cell metabolism
Media optimization.

HOW DO WE APPROACH OMICS DATA?

Observational – "what" is in our cell?

Classical approach

Look at 'natural' diversity in a cell population

Try to reverse engineer what is important

Ex. High vs low producers, omics omics omics

- A very 'normal' PhD

Interventional – how do I introduce meaningful perturbations to understand "why?" and "how?"

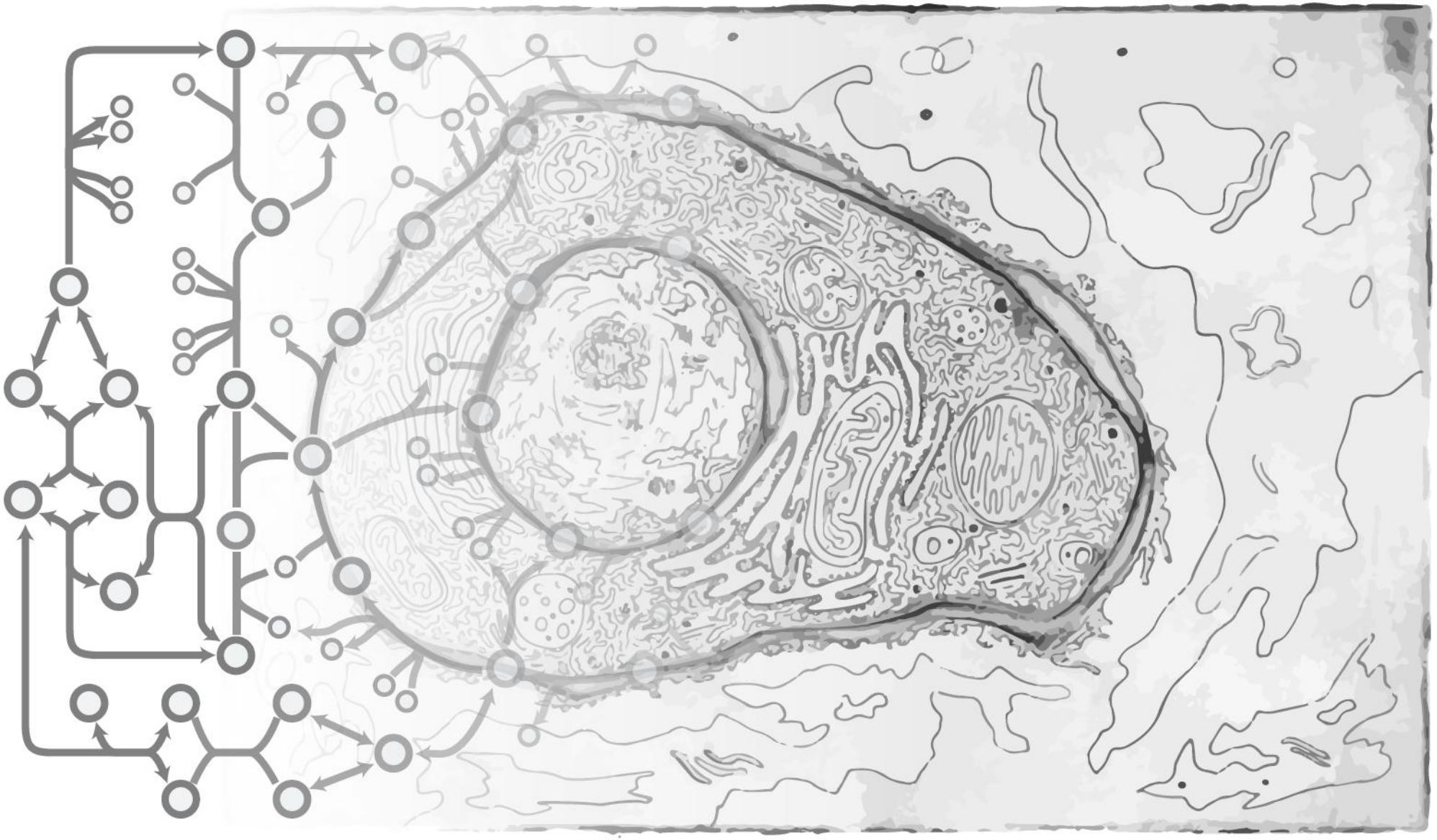
Generate diversity deliberately

Select/identify 'interesting' cells

Have direction genotype/phenotype link

Ex. clinical trials

- No ethical concerns in CHO cell line engineering



CELL PARTS

CLASSES OF BIOLOGICAL MEASUREMENTS

1) Components

- DNA sequence / genotype:
  Next-gen sequencing, SNP & CNV arrays
- Gene expression:
  DNA microarrays, mRNA sequencing
- Protein levels, locations, mods:
  Mass spectrometry, fluorescence microscopy, protein arrays

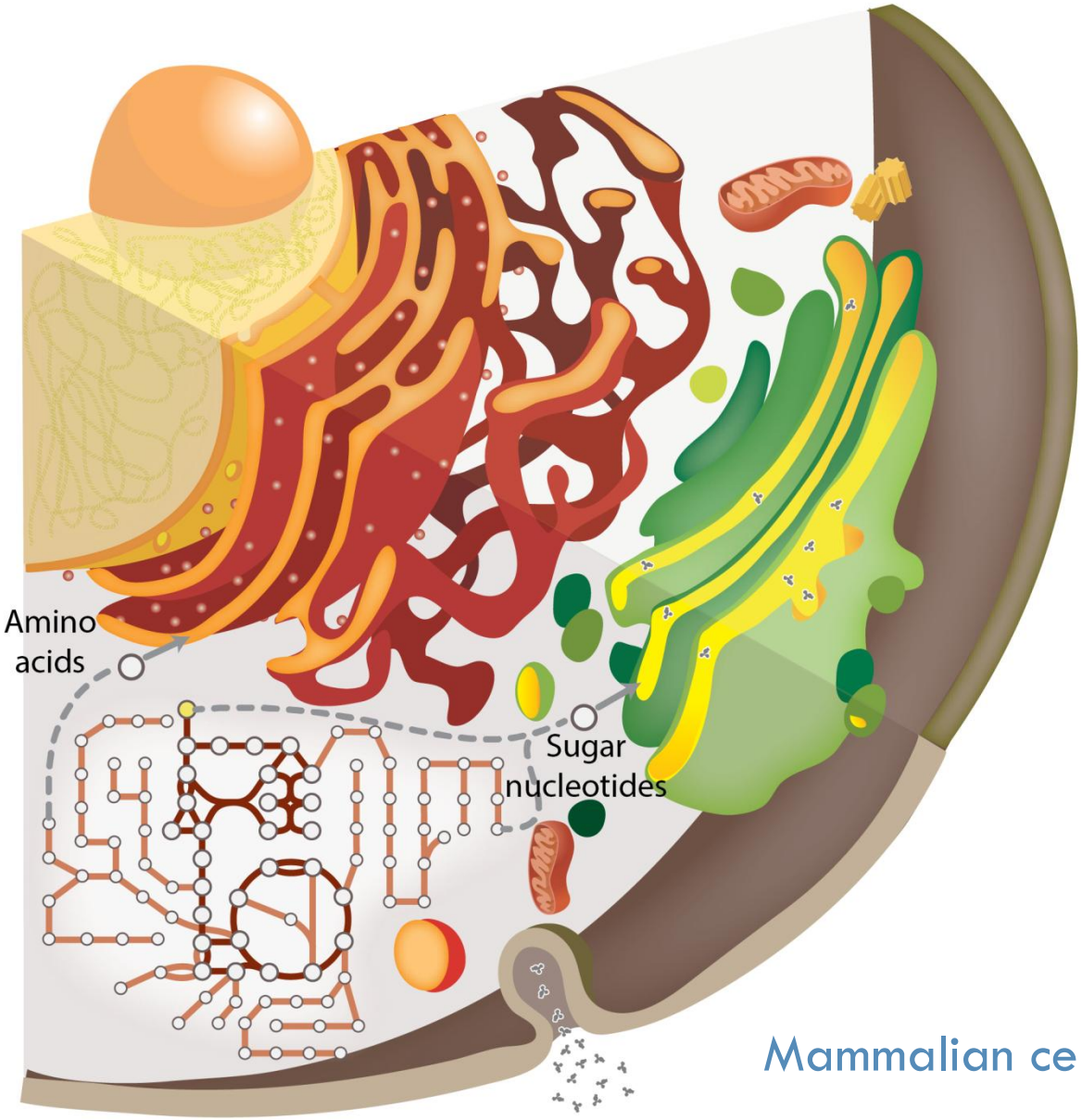
2) Interactions

- Protein-protein interactions:
  Two-hybrid system, coIP, protein antibody array
- Protein-DNA interactions:
  Chromatin IP (chip) sequencing
- Protein-compound

3) Phenotypic traits

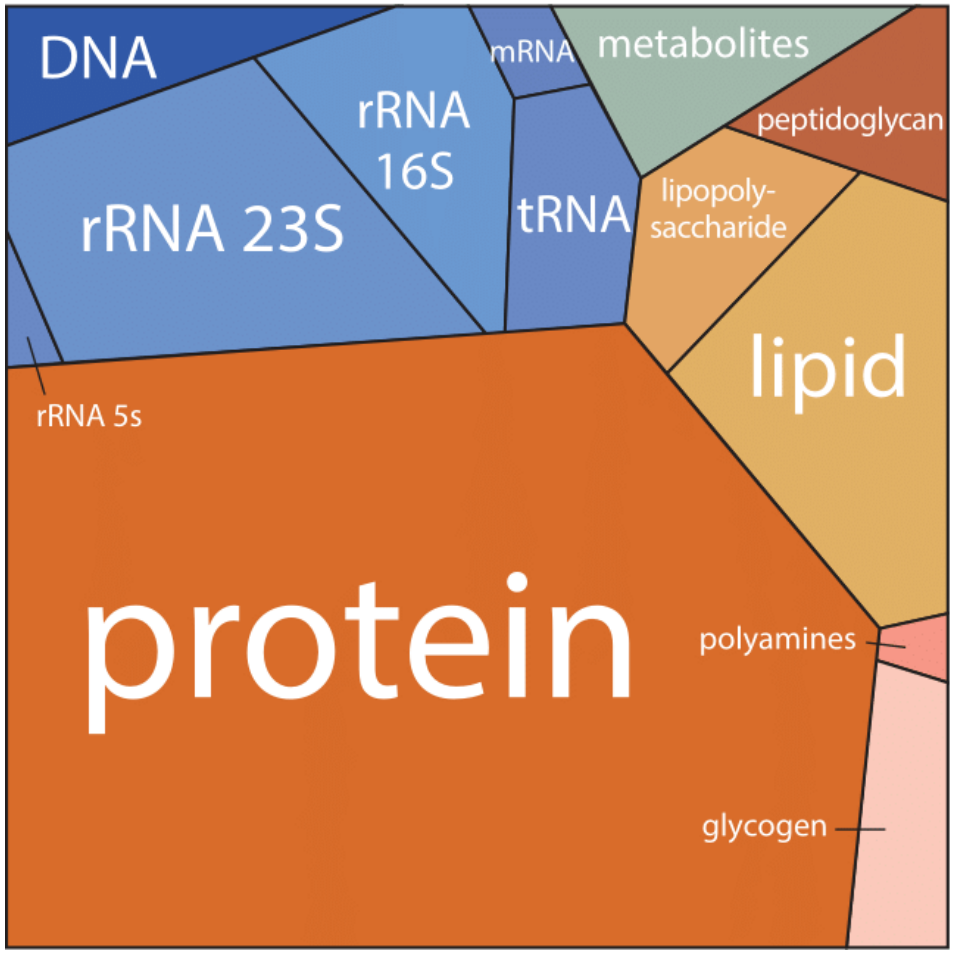
- Physiological or disease state, binary or quantitative
- Growth rate, response to stimulus or stress
- Behaviors

WHAT ARE THE CELL PARTS?

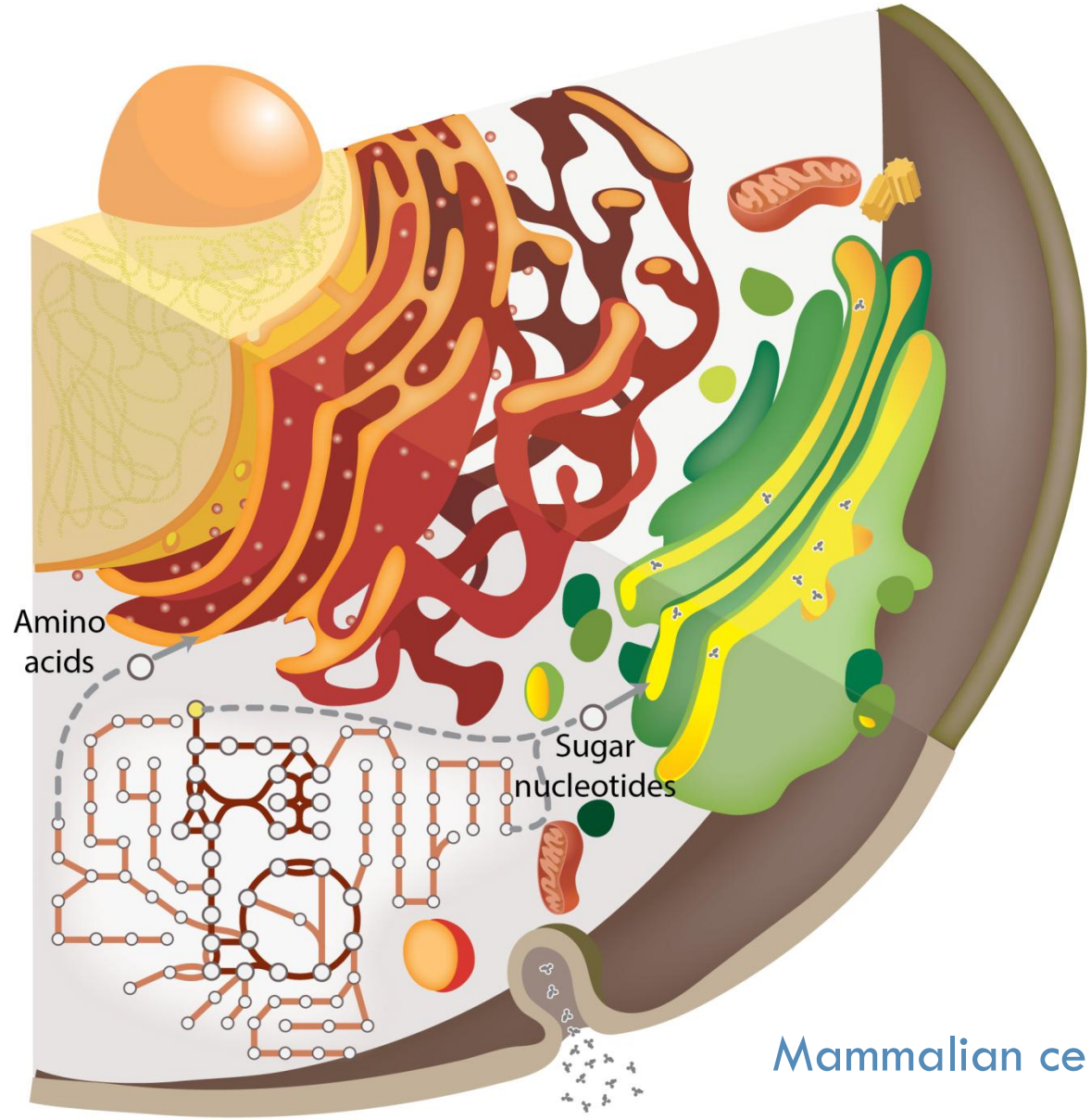


Mammalian cell

E. coli

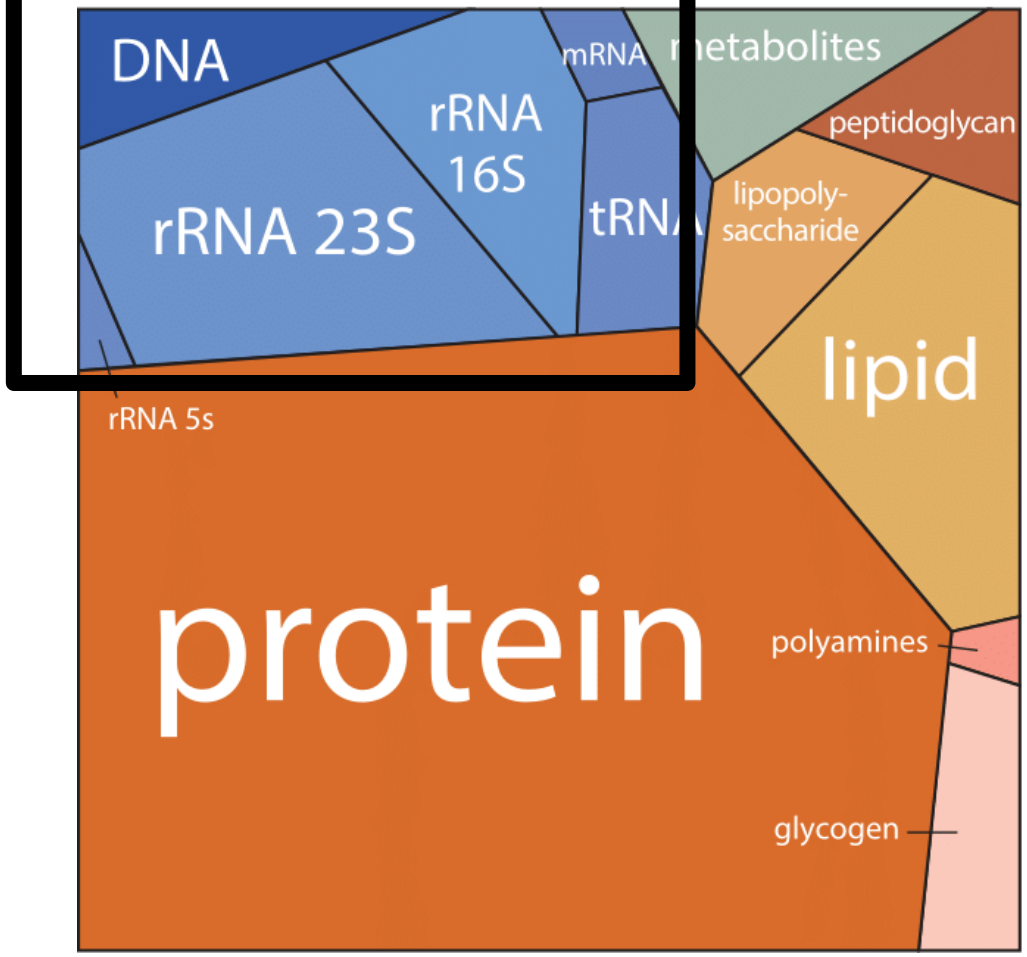


WHAT ARE THE CELL PARTS?



Mammalian cell

E. coli



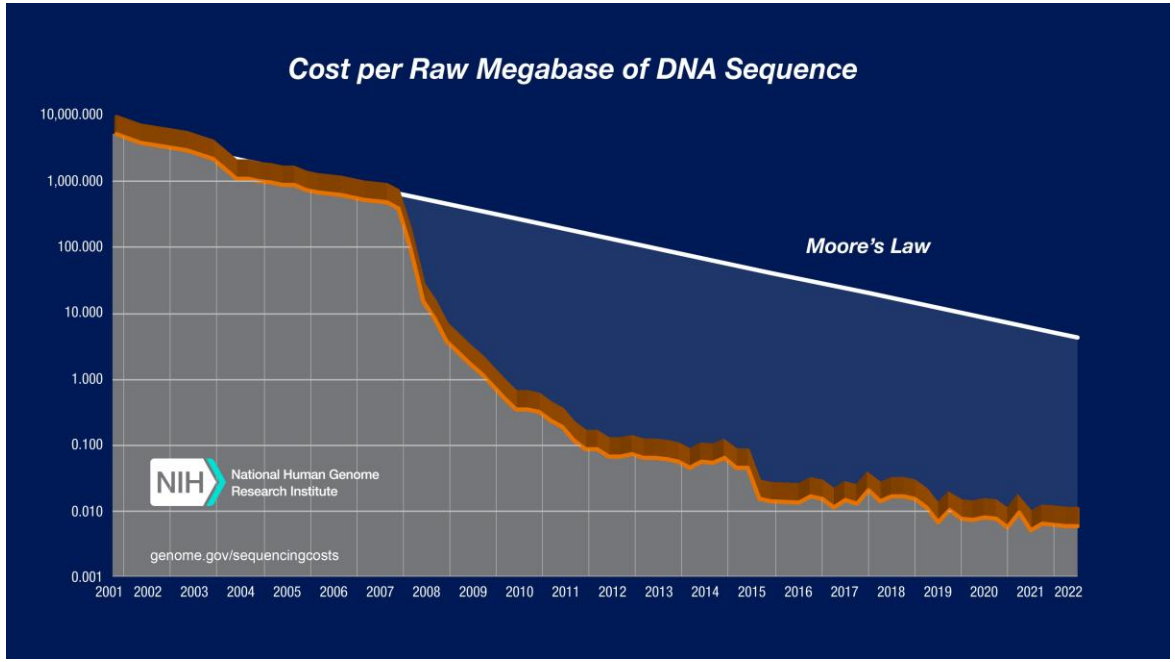
NUCLEIC ACID SEQUENCING

Short read sequencing

Long read sequencing

Pros/cons

What about RNA?



<$1000 human genome currently

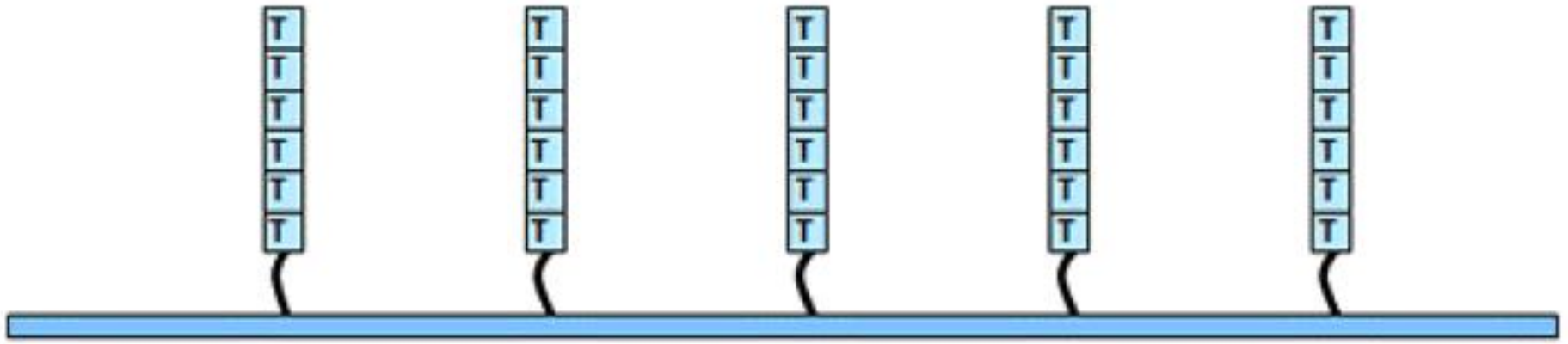
NUCLEIC ACID SEQUENCING: SEQUENCING BY SYNTHESIS WITH REVERSIBLE TERMINATORS

Easiest to see in action…

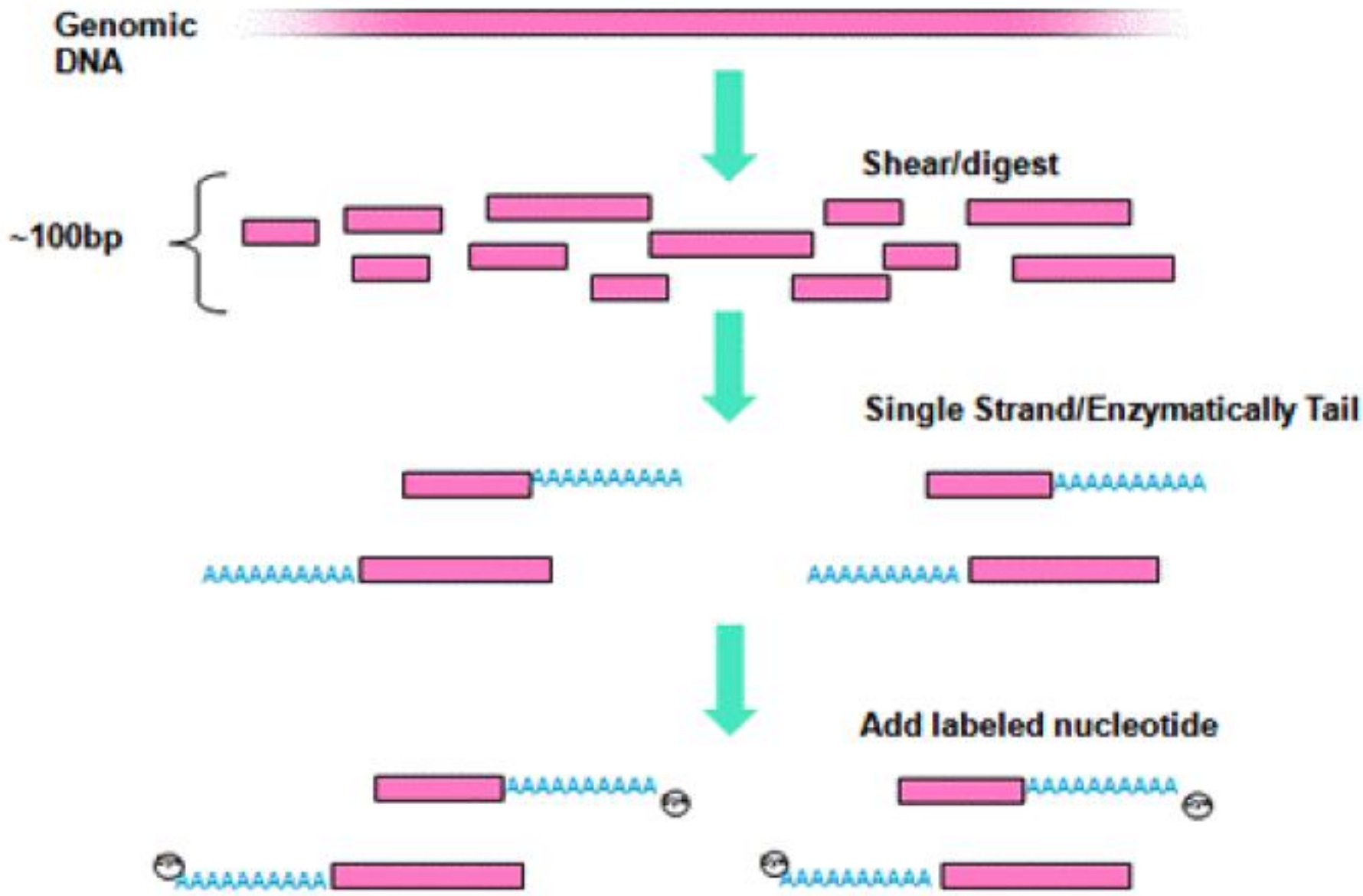
Step 1: Universal primers are immobilized on a glass surface inside a flow cell.

SEQUENCING BY SYNTHESIS

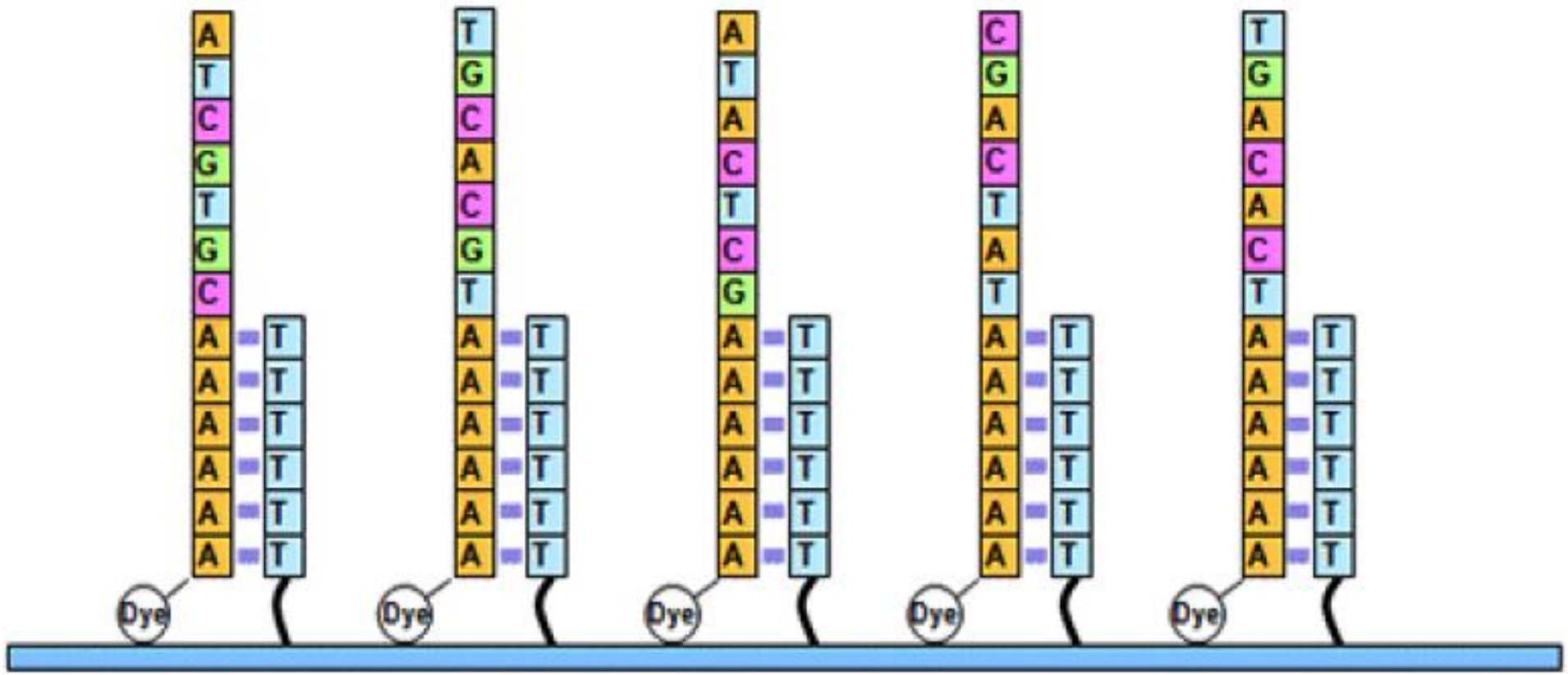
(ILLUMINA GENOMEANALYZER OR HISEQ)



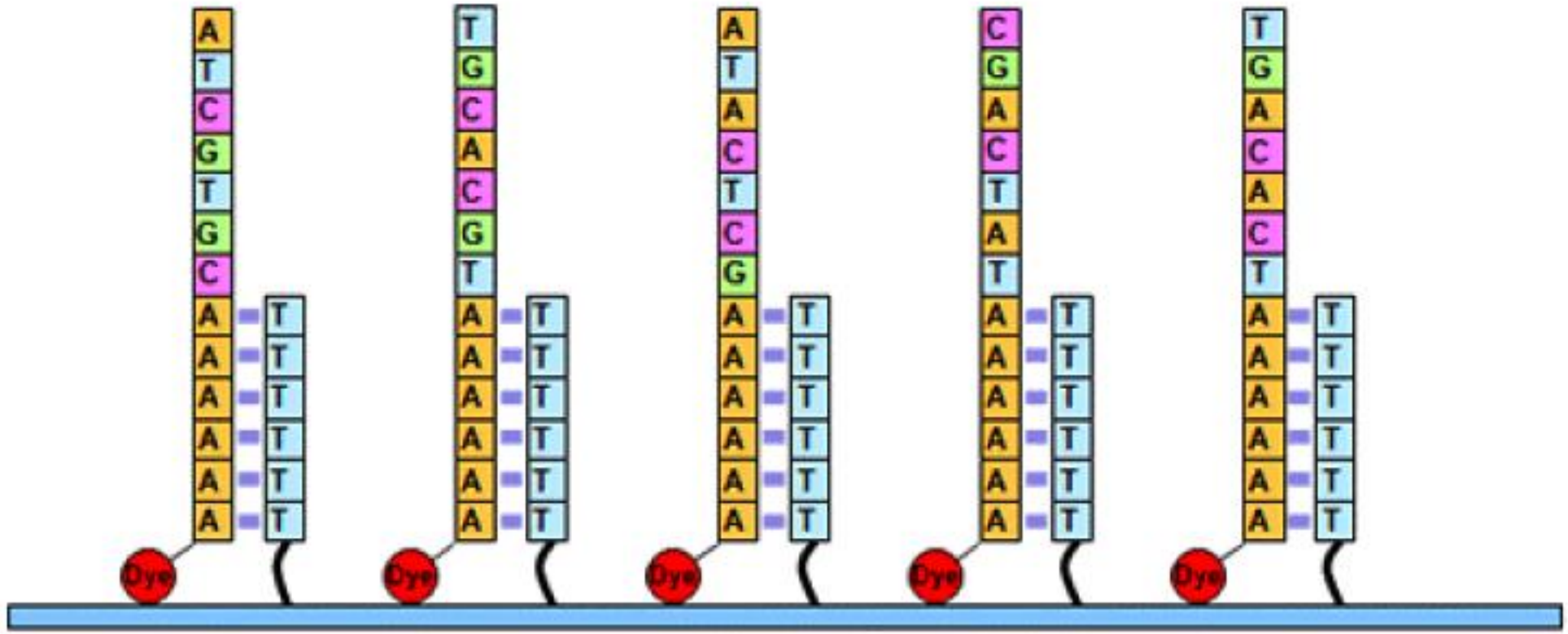
Step 2: Genomic DNA is converted into sequencing templates ready to load into the flow cell.



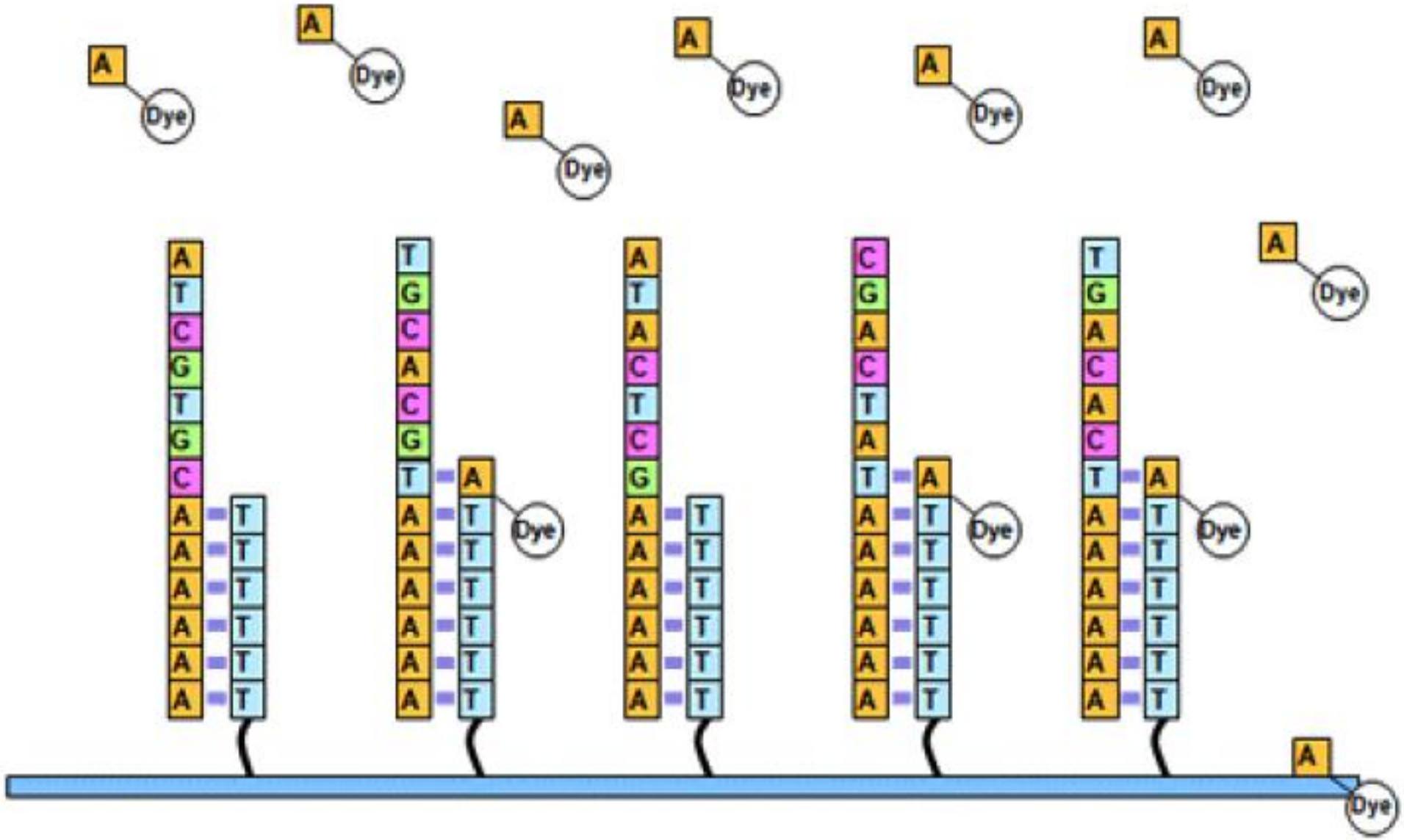
Step 3: Hybridize the DNA templates to the immobilized primers inside the flow cell.



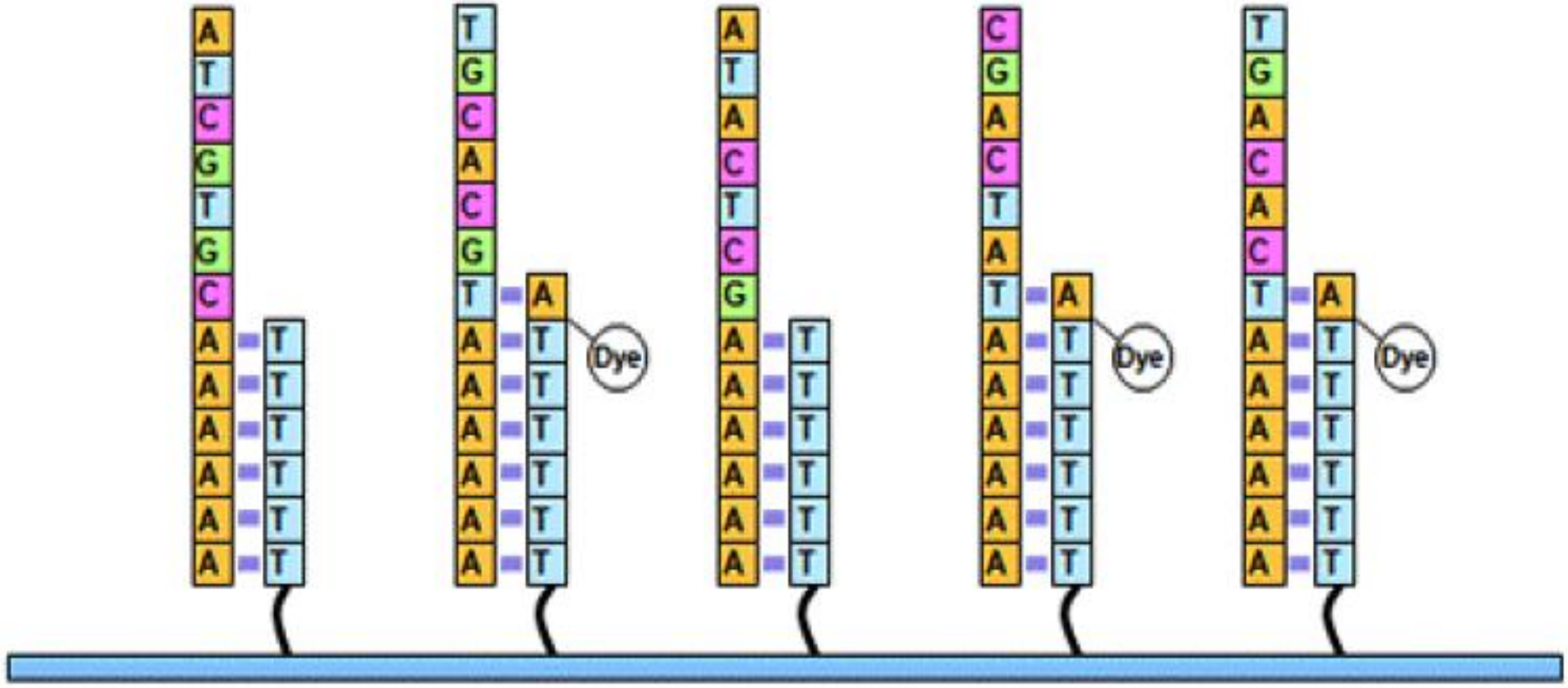
Step 4: Visualize the template:primer duplexes by illuminating the surface with a laser and imaging with an electronic camera connected to a microscope. Record the positions of all the duplexes on the surface. After imaging, the dye molecules are cleaved and washed away.



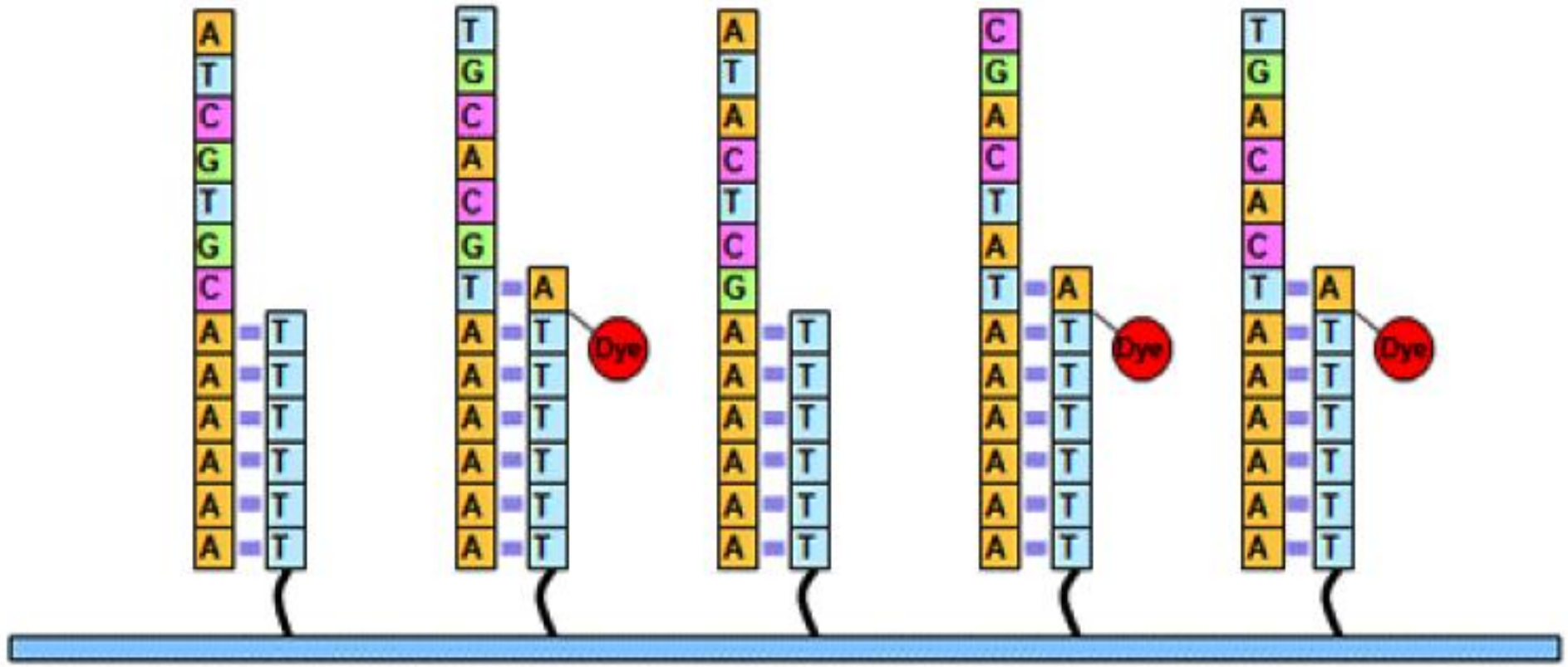
Step 5: Flow in DNA polymerase and one type of fluorescently labeled nucleotide (for example A). The polymerase will catalyze the addition of labeled nucleotide to the appropriate primers.



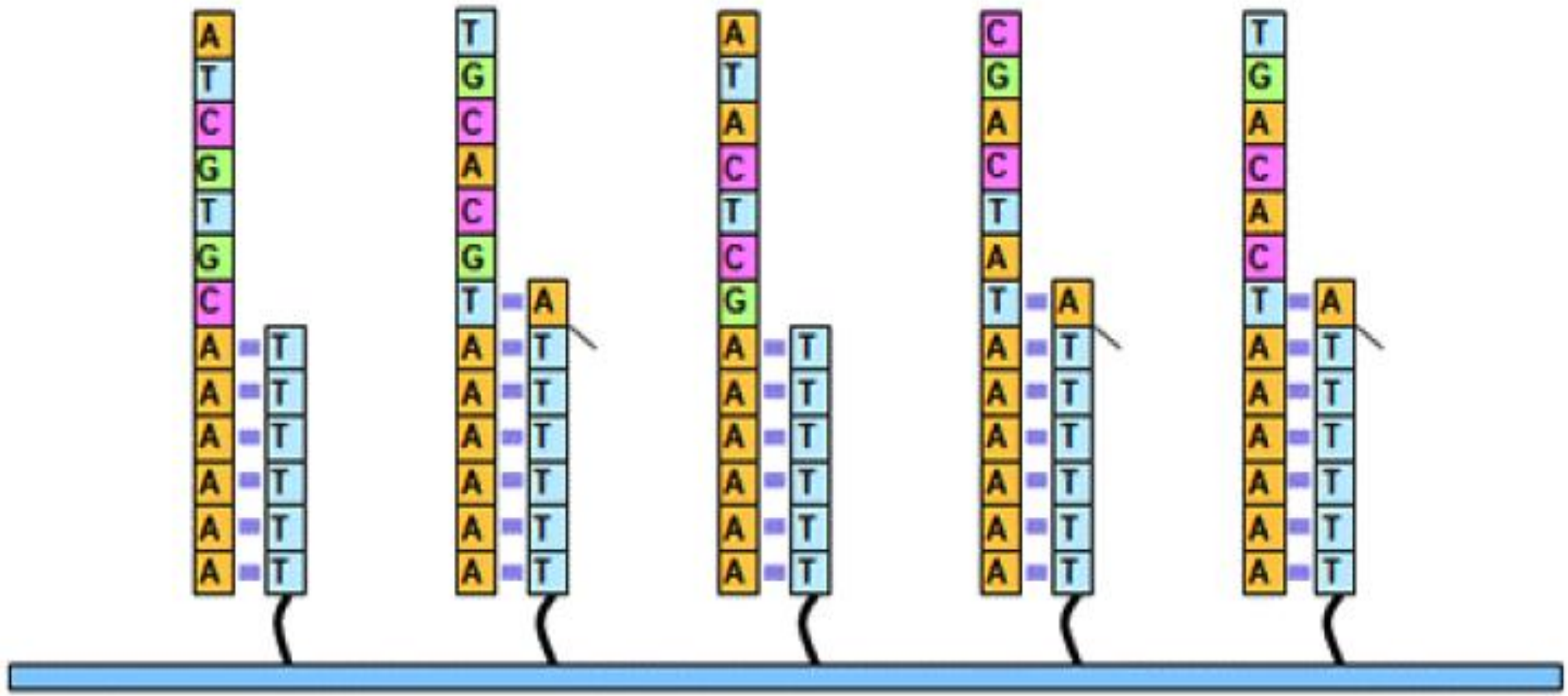
Step 6: Wash out the polymerase and unincorporated nucleotides.



Step 7: Visualize the incorporated labeled nucleotides by illuminating the surface with a laser and imaging with the camera. Record the positions of the incorporated nucleotides.



Step 8: Remove the fluorescent label on each nucleotide.



Step 9: Repeat the process from step 5 with the next nucleotide (stepping through A, C, G and T), until the desired read-length is achieved.

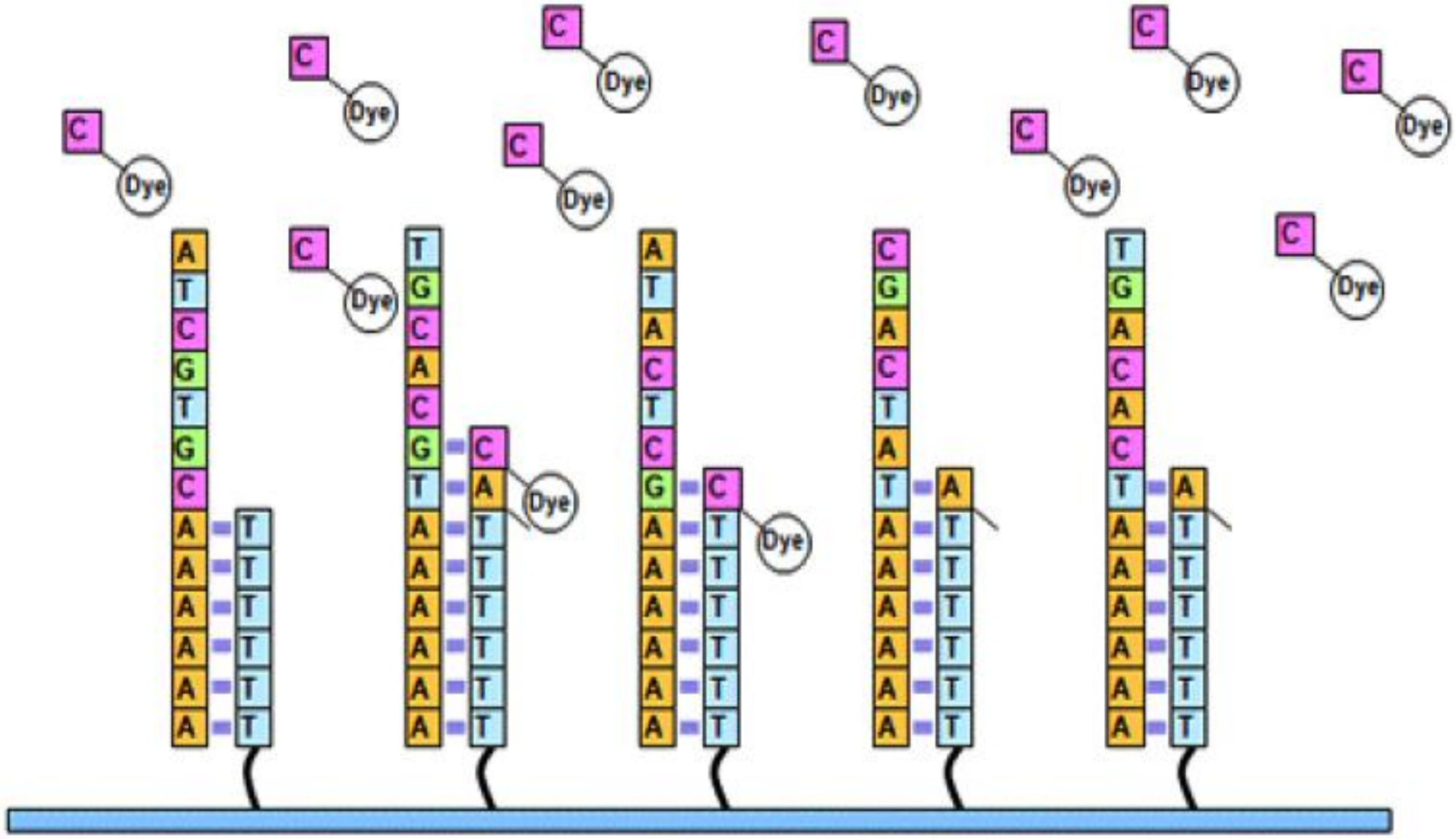
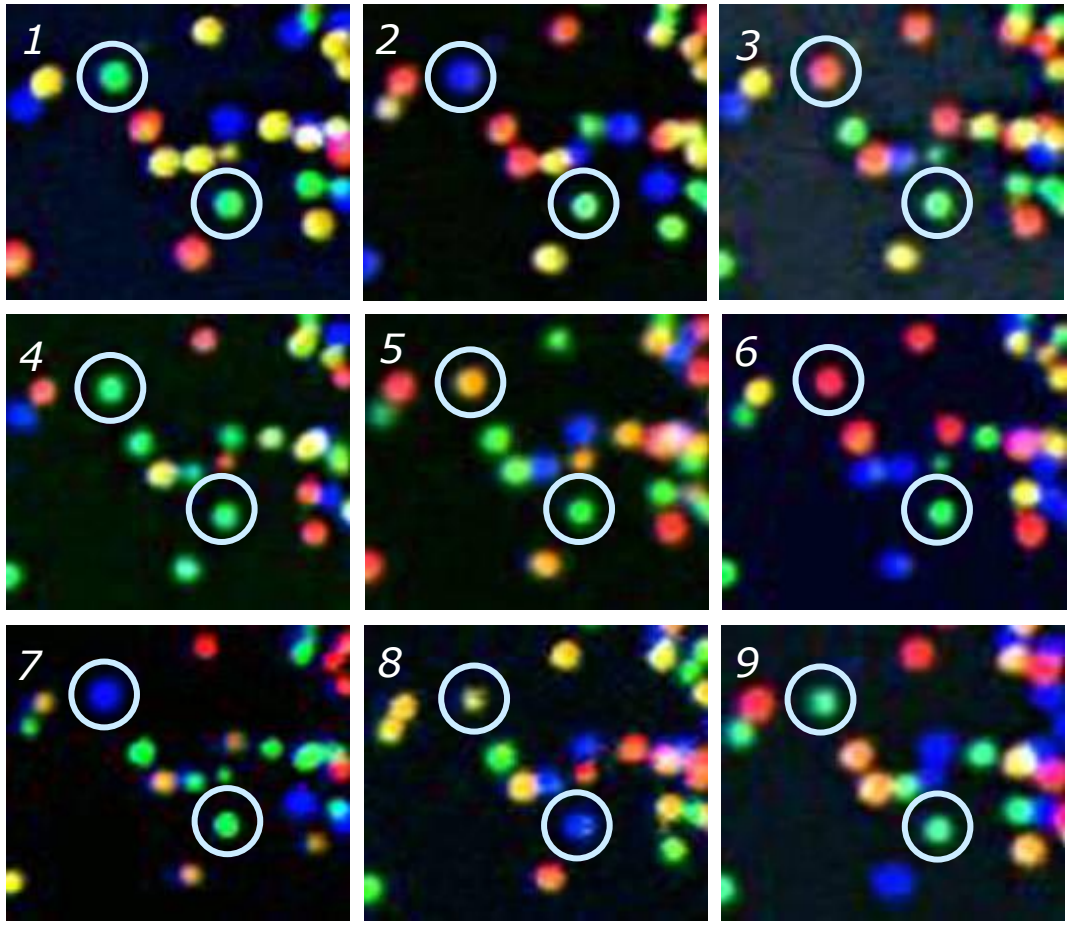
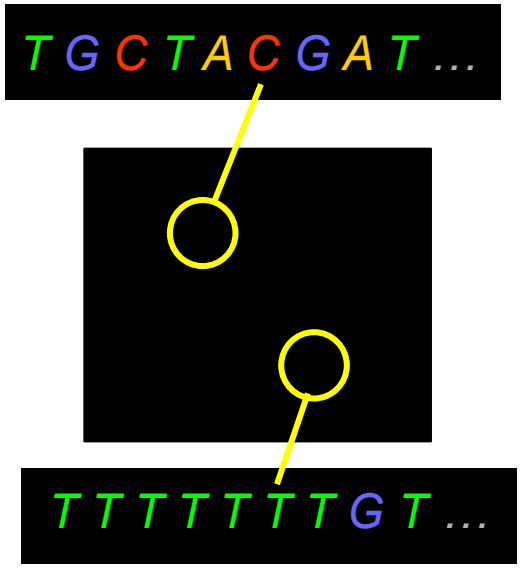
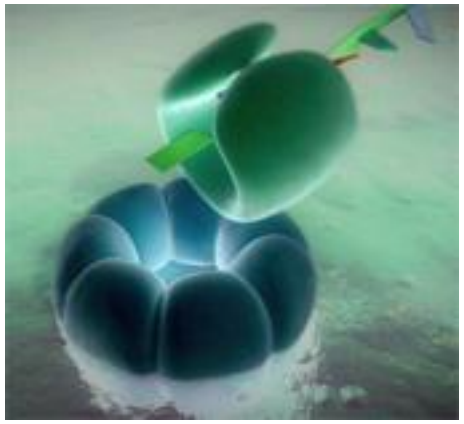


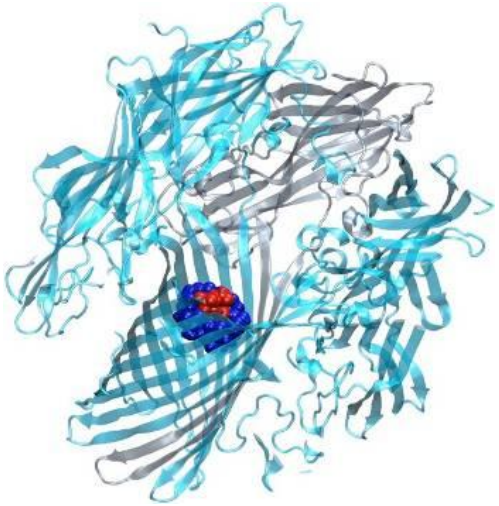
IMAGE TO SEQUENCE





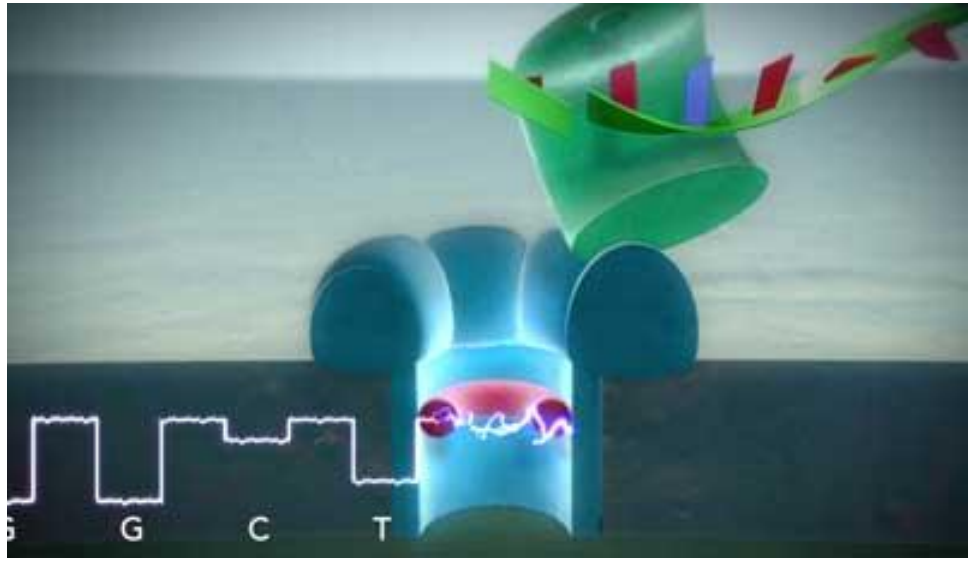
NANOPORE



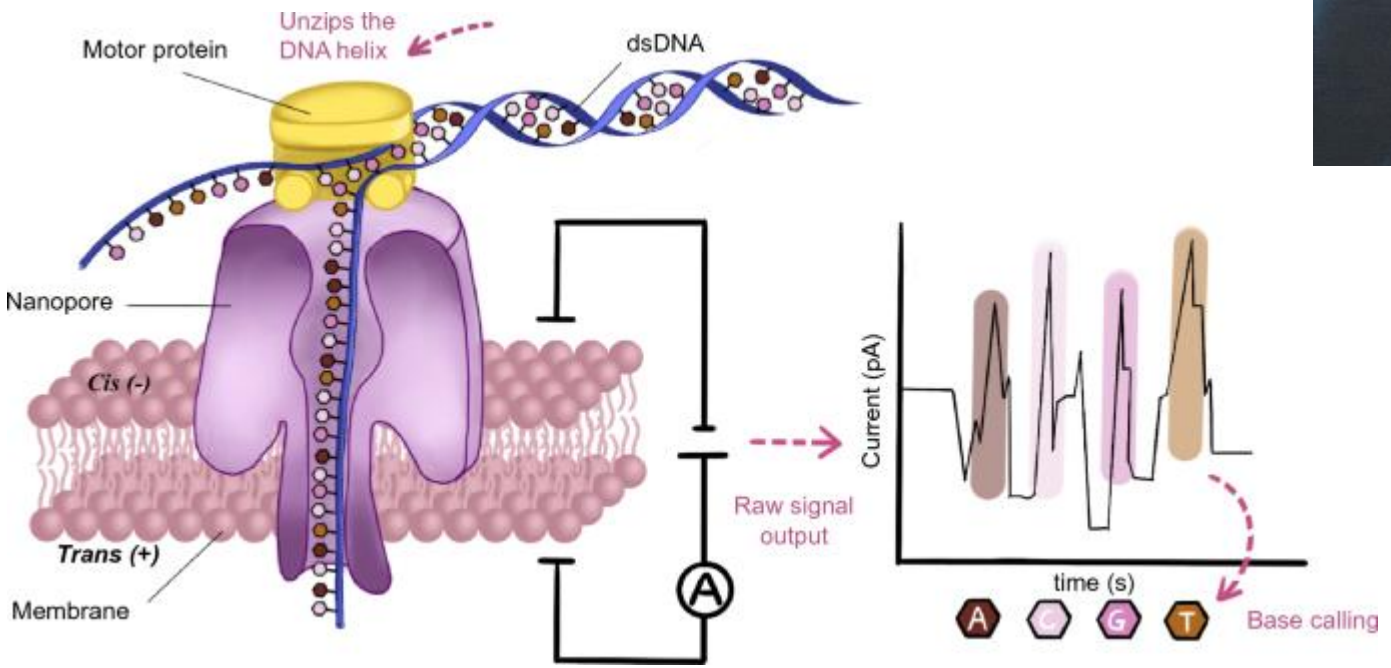


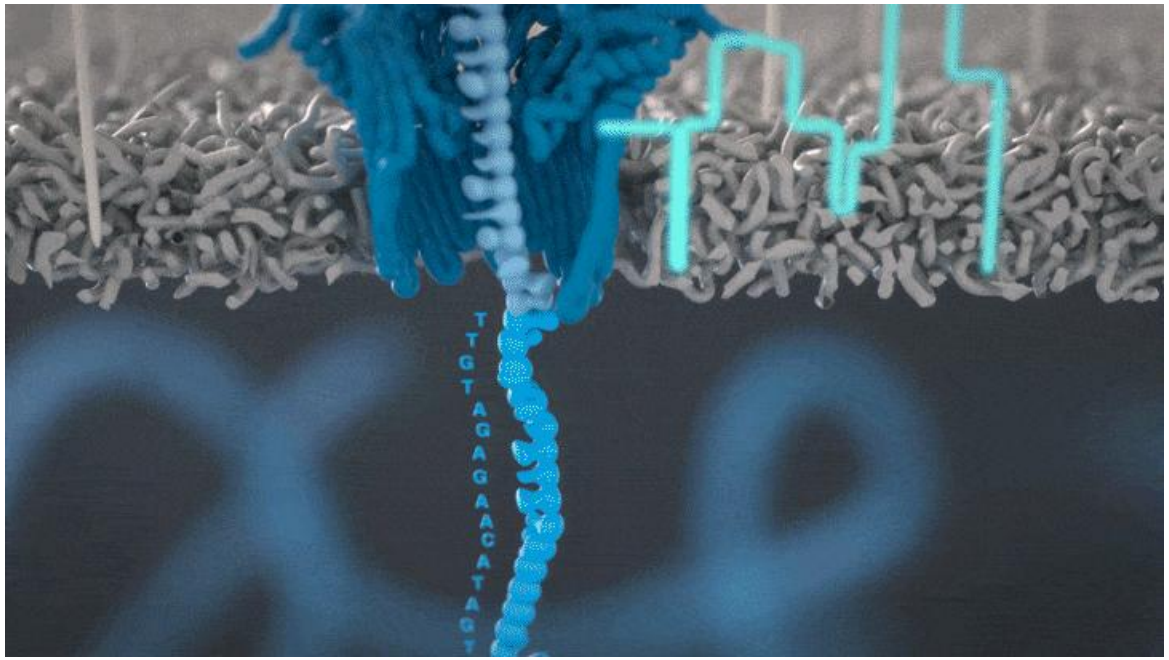
Single molecule sequencing

No labels



IN ACTION



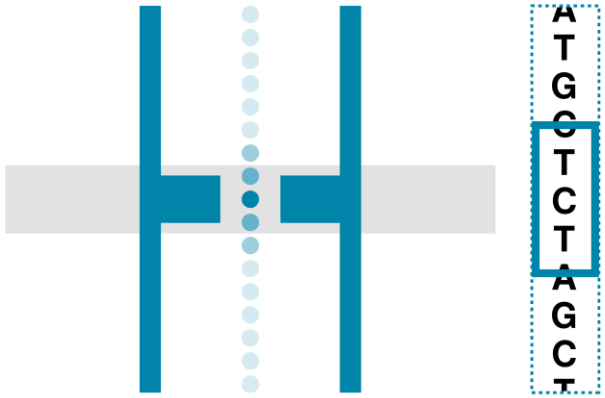


Review > Mol Diagn Ther. 2023 Nov;27(6):685-701. doi: 10.1007/s40291-023-00669-8. Epub 2023 Aug 11.

Application of Nanopore Sequencing in the Diagnosis and Treatment of Pulmonary Infections

Jie Chen [1], Feng Xu [2]

UNDER THE HOOD



Signal is a function of 5 nucleotides

→>1024 “signals” (modified nucleotides)

Noisy signal

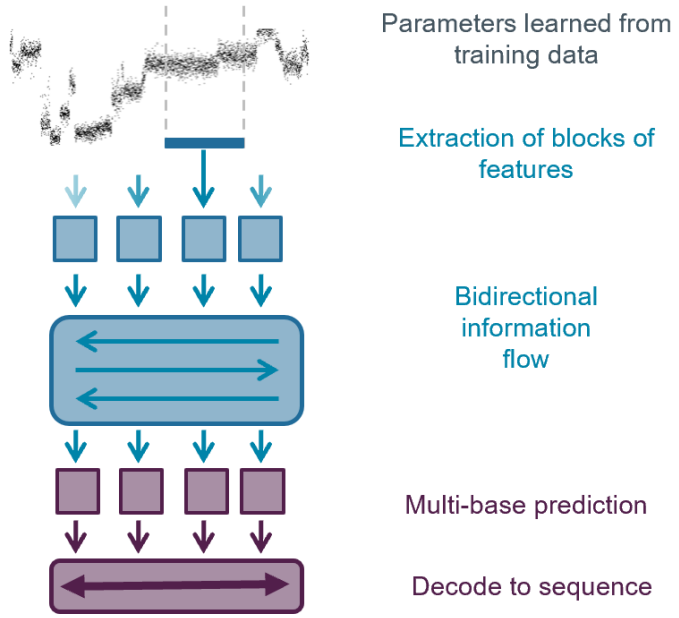
Translocation speed not constant

→difficulty with homopolymer

New pore design

Less difficulty with homopolymers

Greater signal complexity



PROS/CONS

SHORT READ

Cheaper (per base)

Higher throughput (bases sequenced)

(Generally) more accurate

LONG READ

Can capture large SVs

Spans repetitive regions

Powerful for de novo genome assembly

May be able to directly sequence modifications

Hybrid approaches growing in popularity

WHAT ABOUT RNA?

Not stable

BMC Genomics. 2021 Jan 21;22(1):69. doi: 10.1186/s12864-021-07381-z.

Multiple freeze-thaw cycles lead to a loss of consistency in poly(A)-enriched RNA sequencing

Benjamin P Kellman [# 1 2], Hratch M Baghdassarian [# 1 2], Tiziano Pramparo [3], Isaac Shamie [1 2], Vahid Gazestani [1 3], Arjana Begzati [4], Shangzhong Li [1 5], Srinivasa Nalabolu [3], Sarah Murray [6], Linda Lopez [3], Karen Pierce [3], Eric Courchesne [3], Nathan E Lewis [7 8 9]

→convert to DNA* via RT

* Can directly sequence on Nanopore

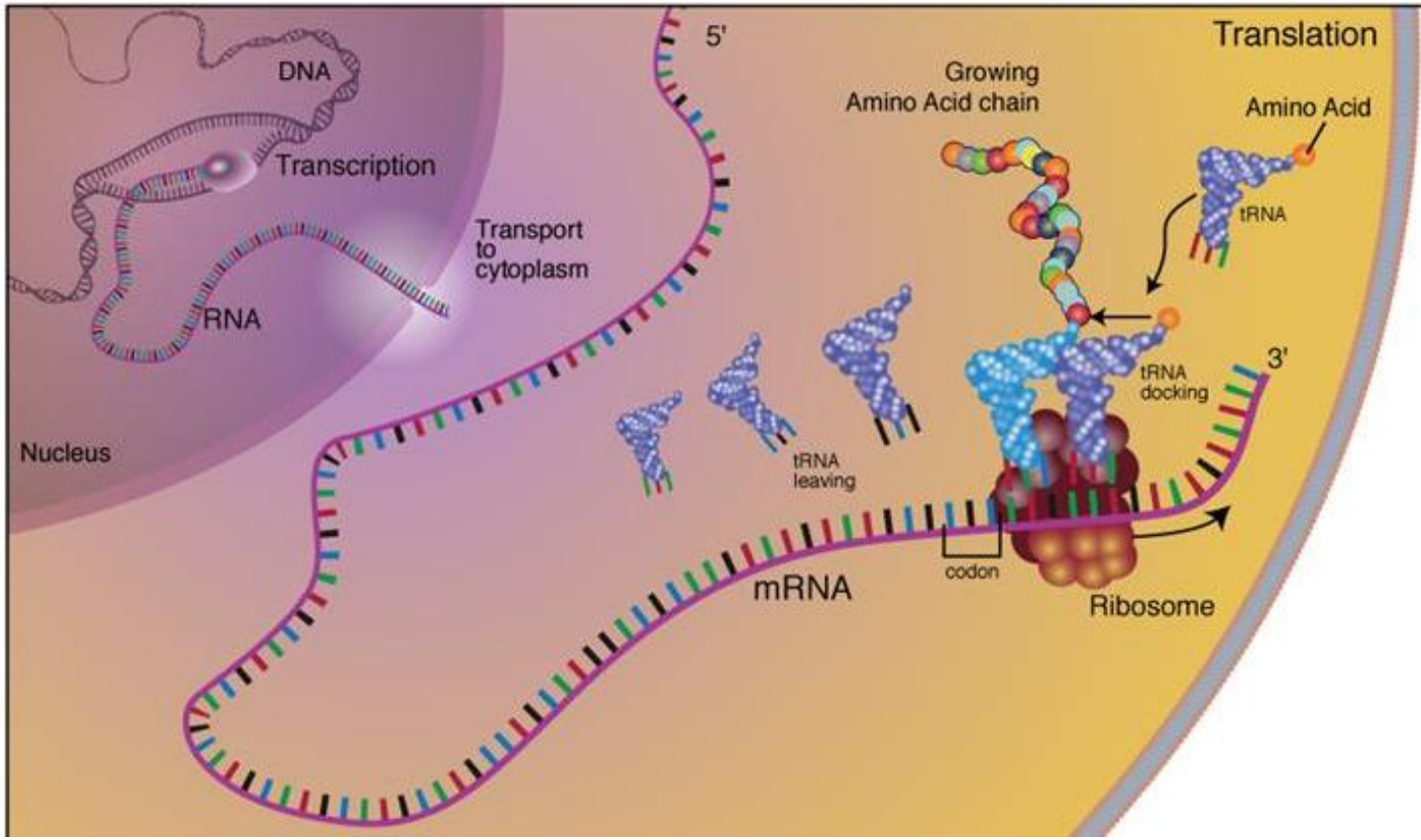
How to deal with rRNA? (90% of RNA!)

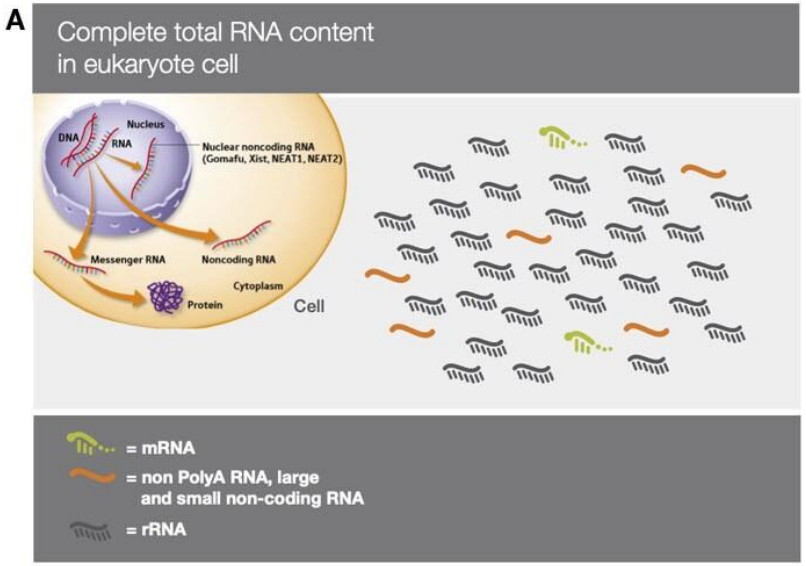
→rRNA depletion

→polyA enrichment

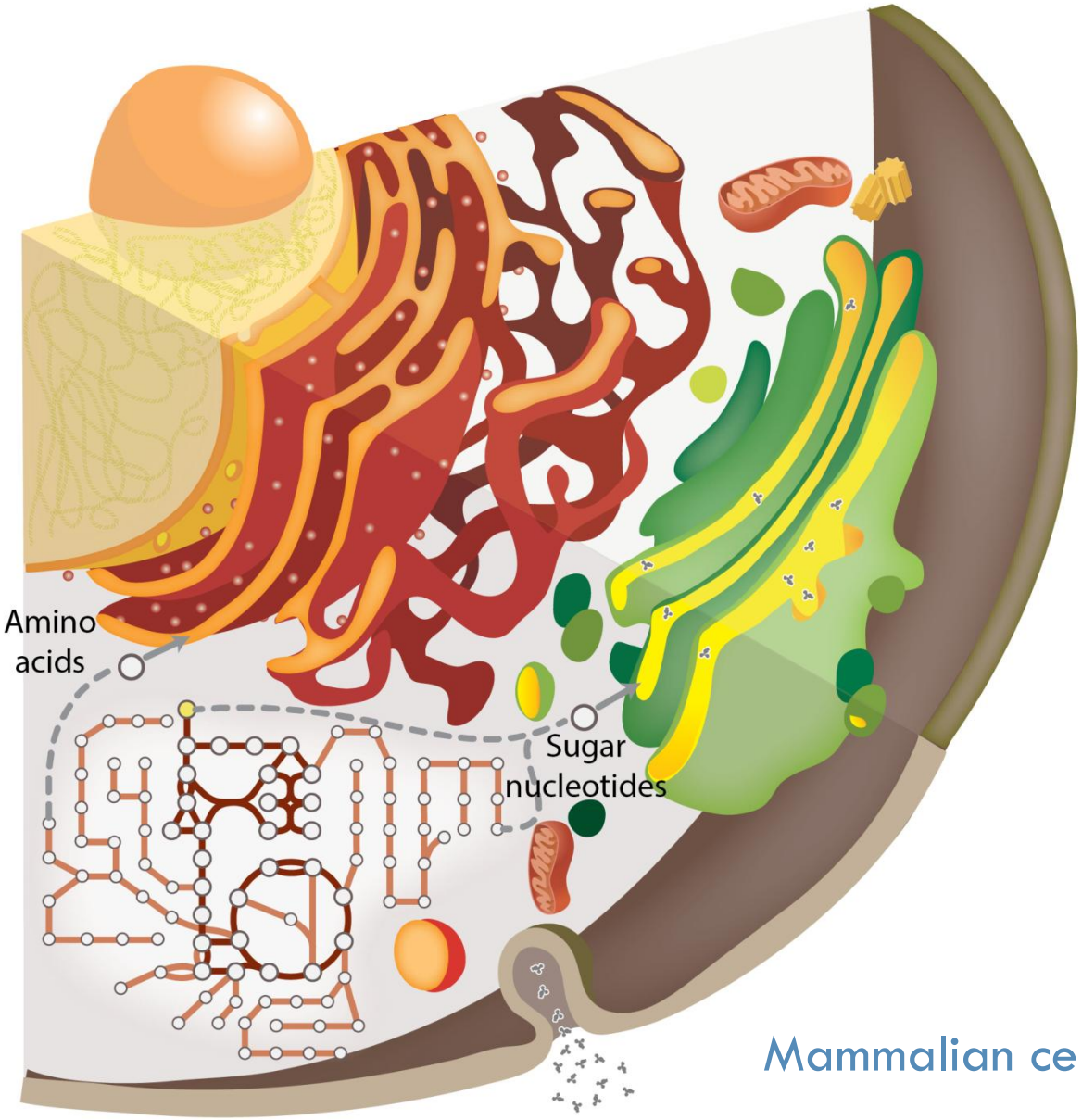
Can also directly sequence w/long read

→isoform discovery



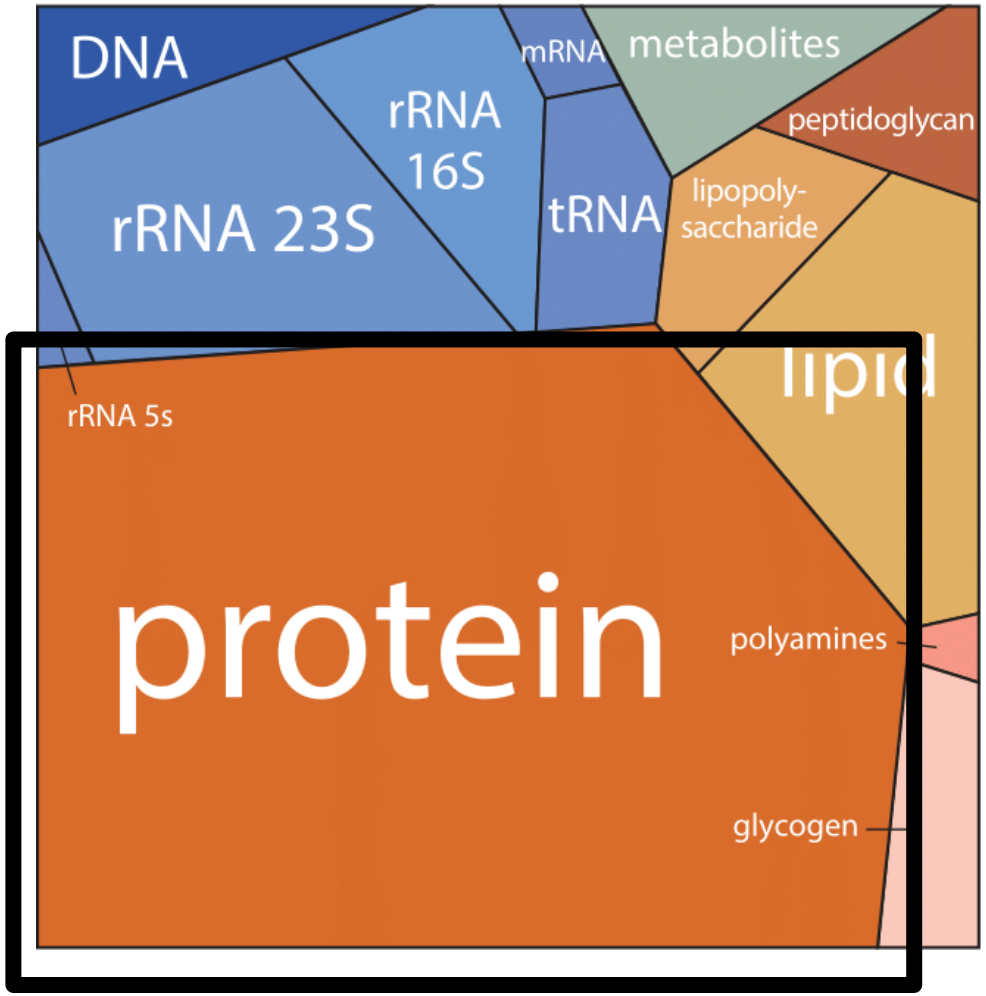


WHAT ARE THE CELL PARTS?



Mammalian cell

E. coli



WHAT ARE PROTEOMICS?

Protein

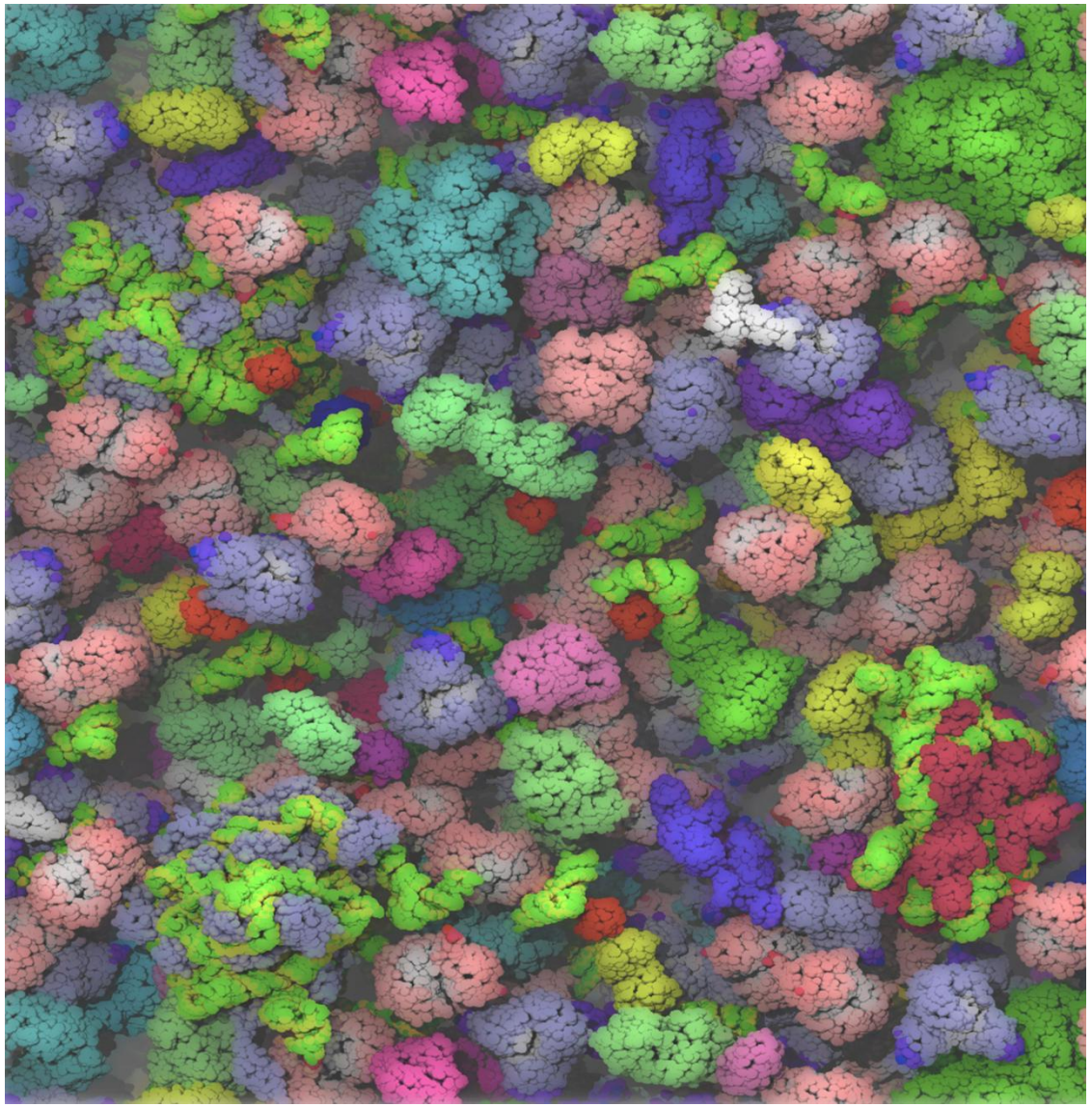
- Identification
- quantification
- Characterization

Large-scale study of protein

- Expression
- Regulation
- Modification
- Function
- Interactions

Proteomic tools

- Separation approaches
- Mass spectrometry
- Microscopy and advanced imaging



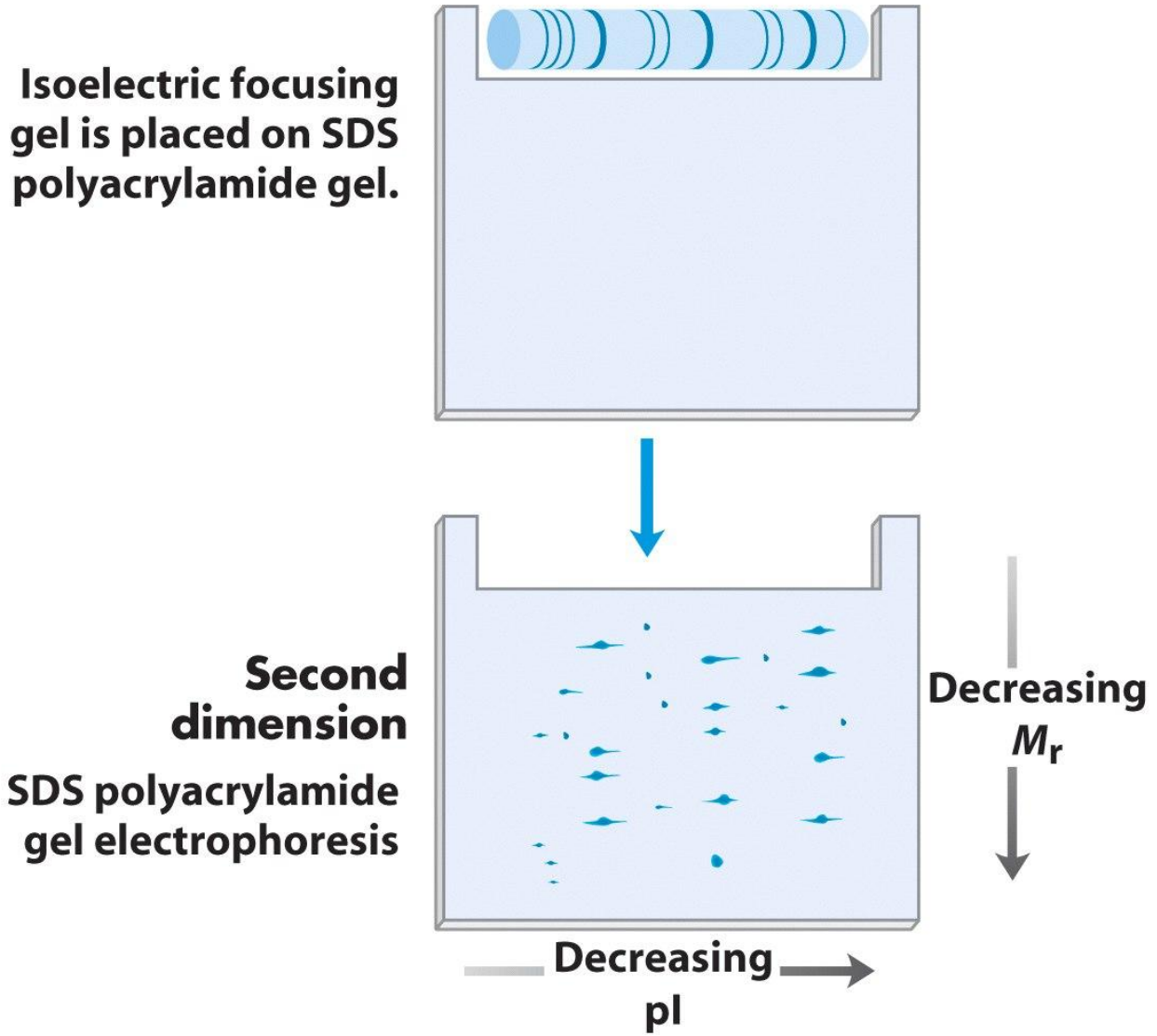
HOW DO WE MEASURE PROTEINS?

Electrophoresis

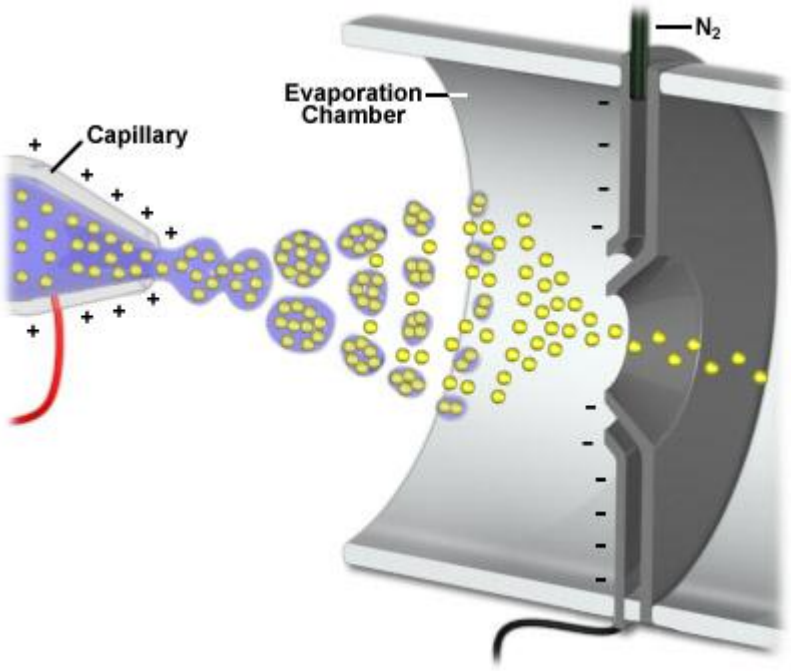
Mass spectrometry

- Ionization
  - MALDI - Matrix-assisted laser desorption/ionization
  - ESI – electrospray ionization
- Mass measurement
  - TOF - Time Of Flight
  - Quadrupole

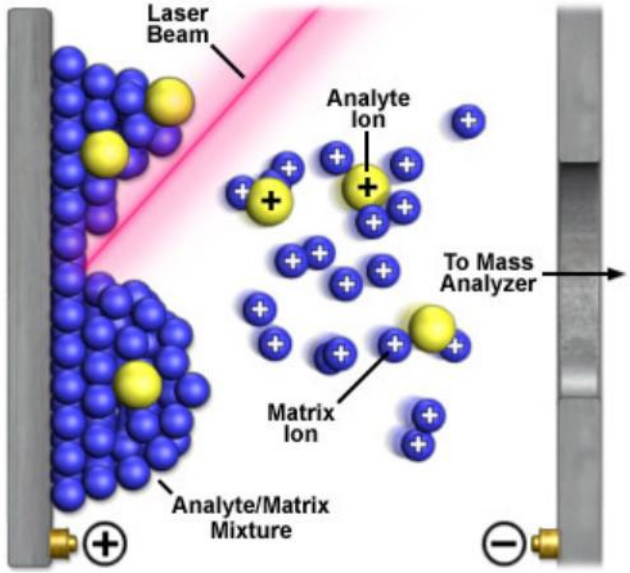
2-D GEL ELECTROPHORESIS



IONIZATION METHODS FOR MASS SPECTROMETRY



Electrospray ionization (ESI)



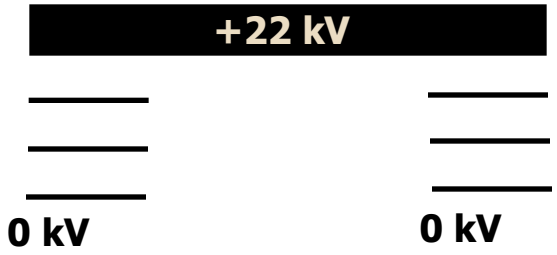
Ionization of matrix and sample particles as a result of laser exposure.

Image used with permission from Dr. Chris Hendrickson, Florida State University (2).

Matrix assisted laser desorption ionization (MALDI)

MASS SPECTROMETRY - TOF

Time-of-Flight MS

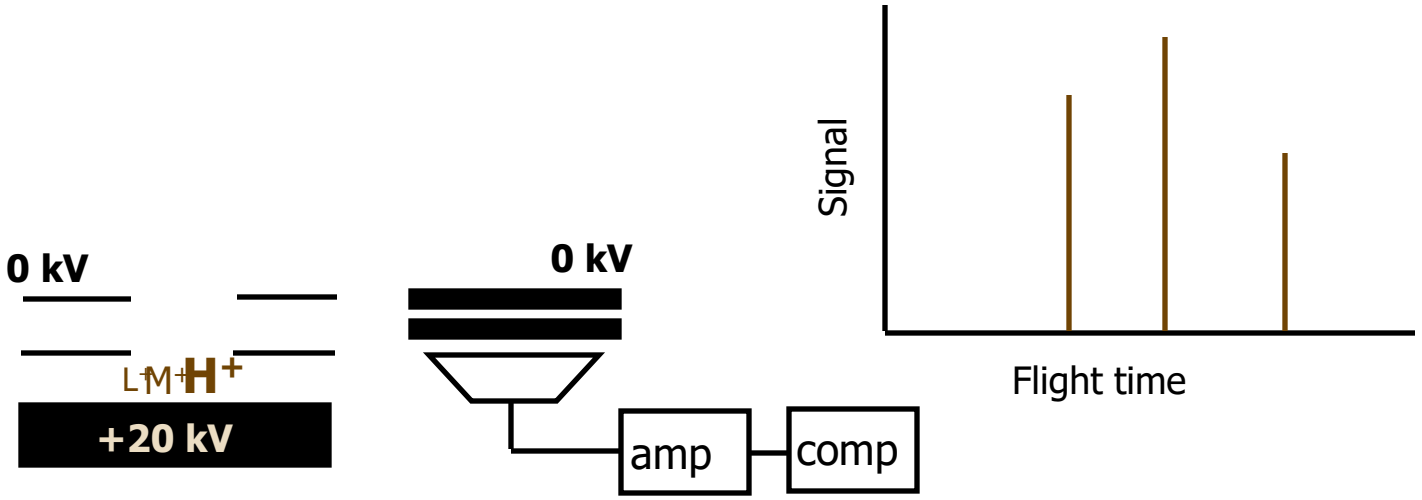


1) Ions enter source region, accelerated toward reflectron.

2) Ions separate in space based on their relative mass-to-charge (m/z).

3) Ions reverse path in reflectron.

4) Ions impact detector.



MASS SPECTROMETRY – QUADRUPOLE MASS ANALYZER

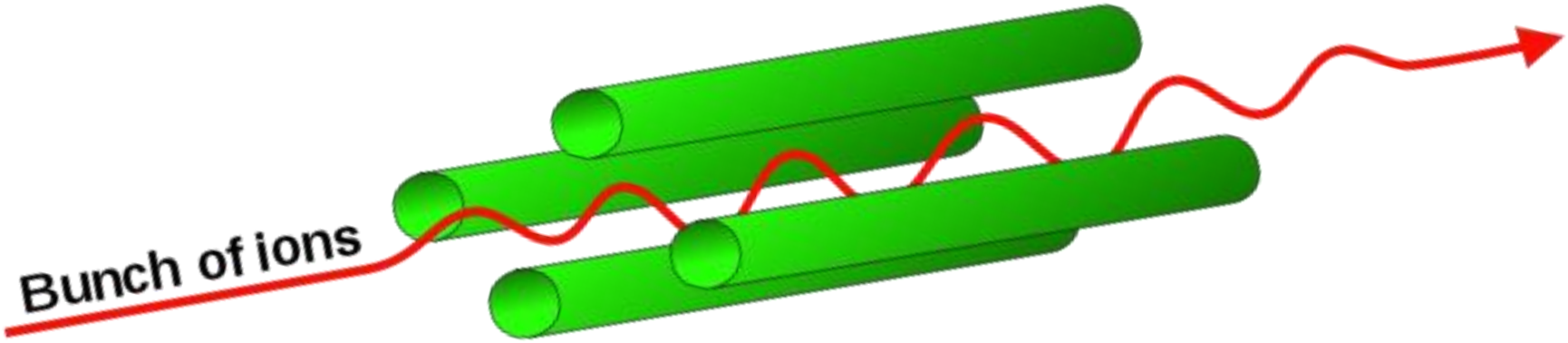
Four parallel metal rods

Opposing rod pair is connected together electrically

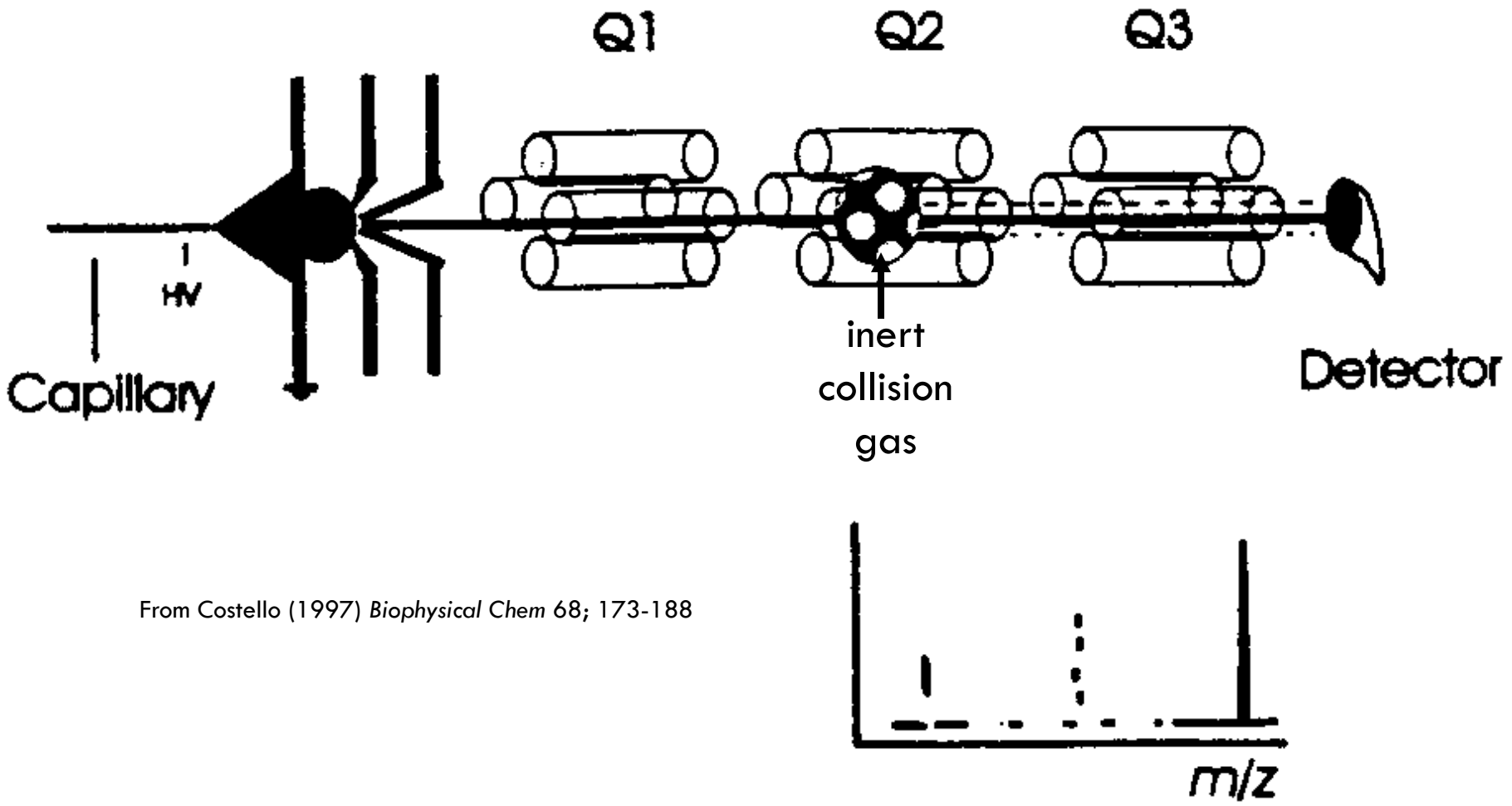
A radio frequency (RF) voltage applied between one pair of rods and the other

DC voltage is superimposed on the RF voltage

Ions traveling down the quadrupole between rods are filtered based on m/z

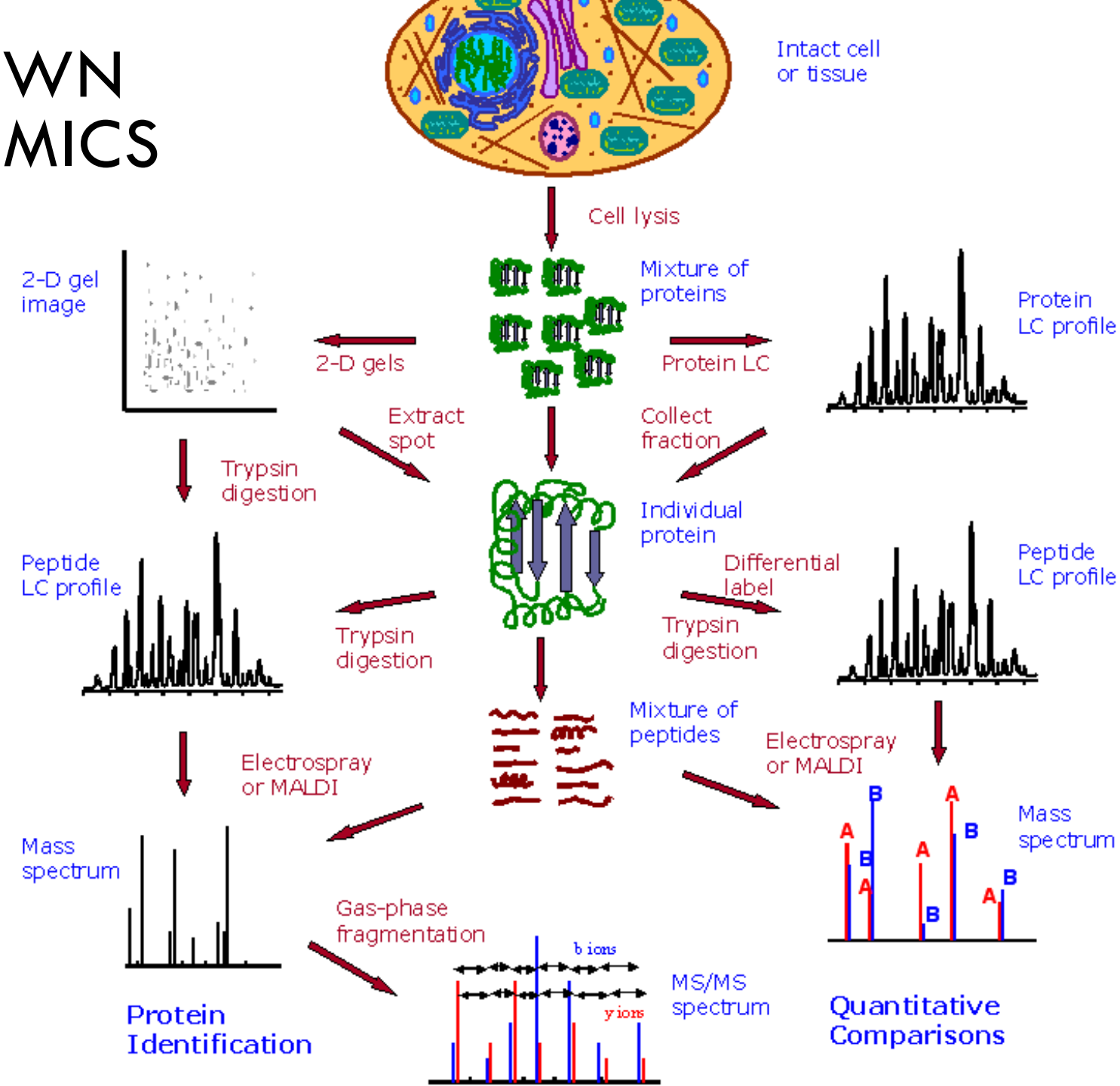


MASS SPECTROMETRY – TRIPLE QUADRUPOLE

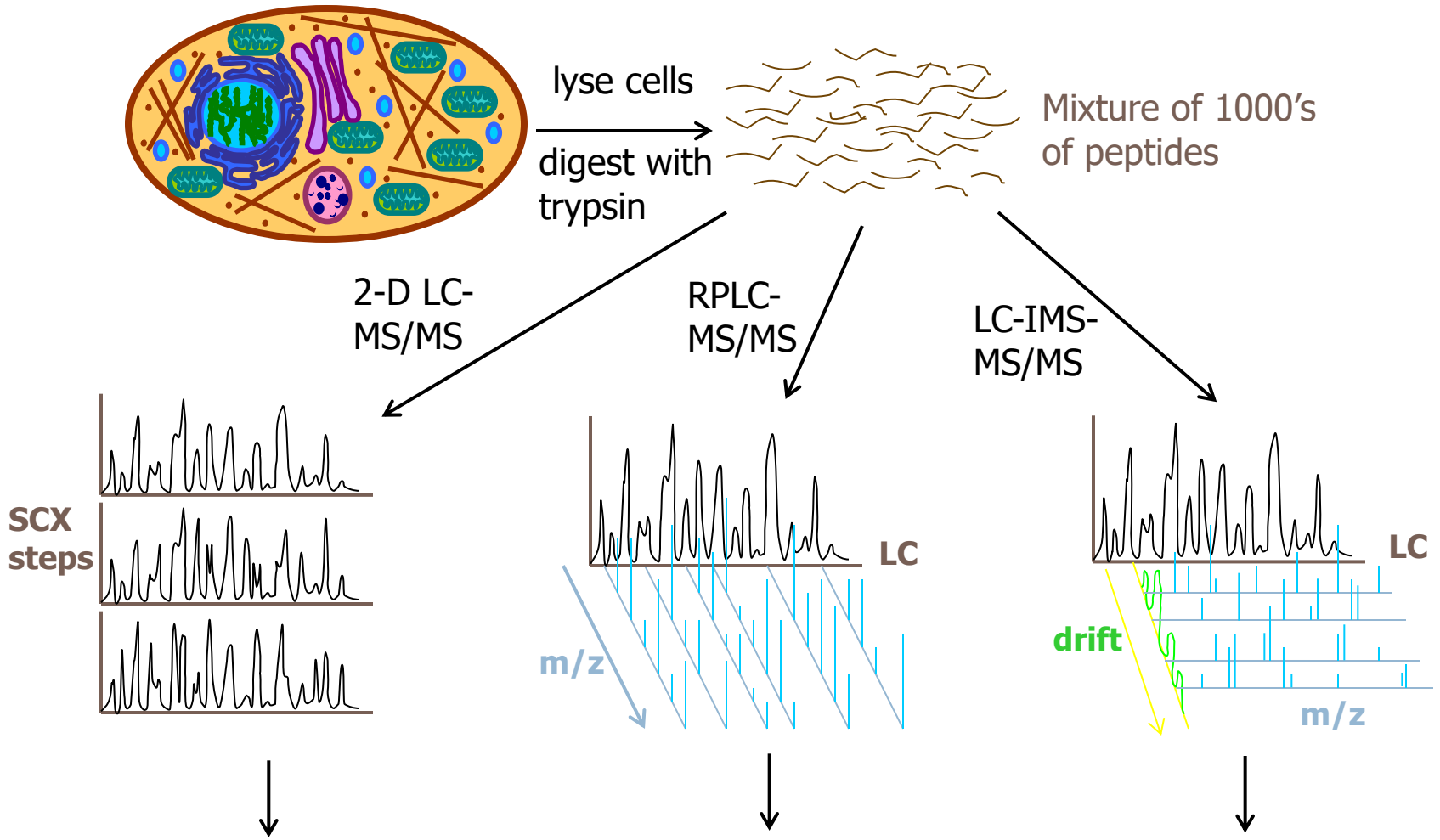


From Costello (1997) *Biophysical Chem* 68; 173-188

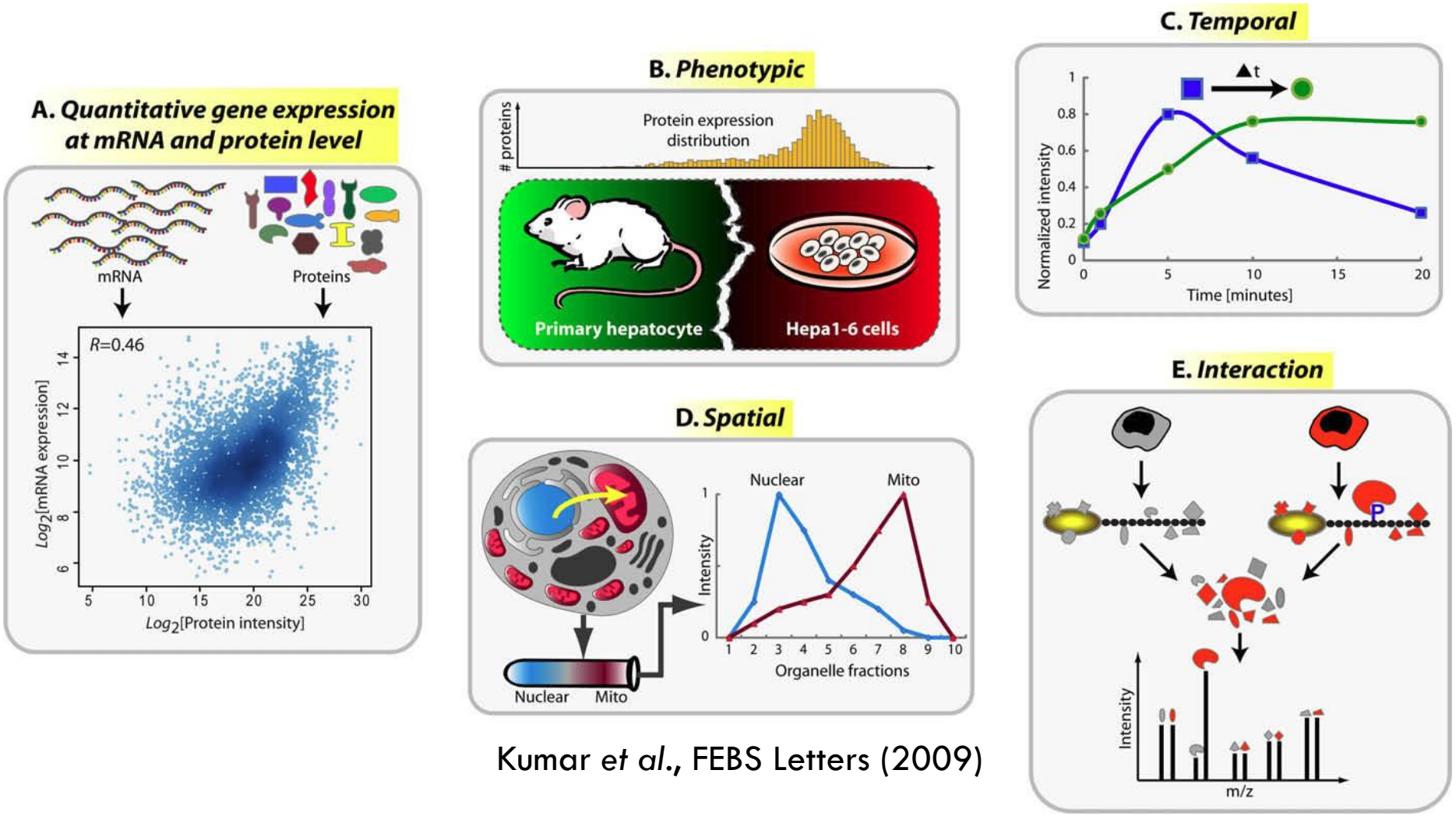
TOP-DOWN PROTEOMICS



"SHOTGUN" PROTEOMICS

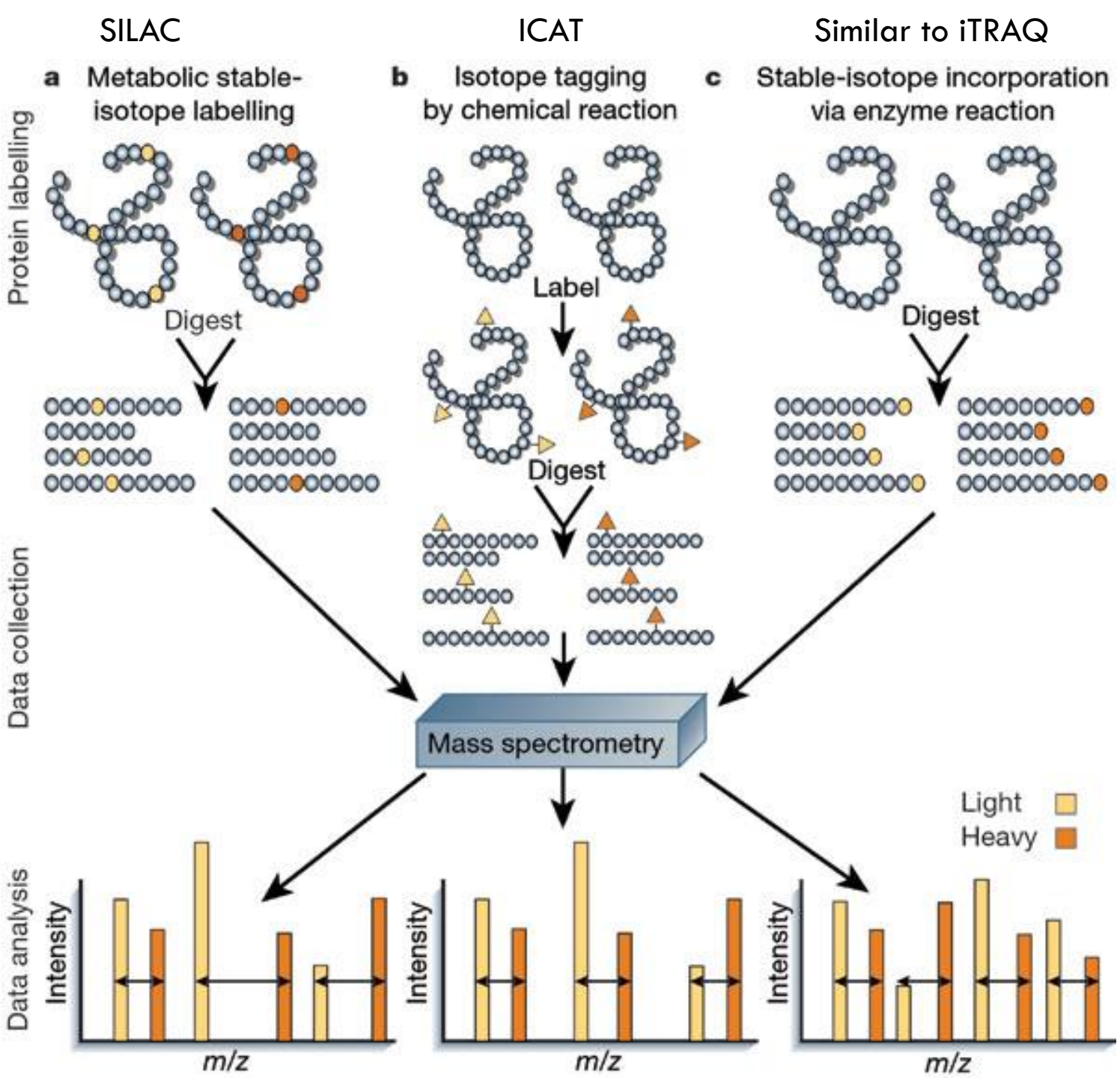


Knowing what my protein is is nice… but how much do I have?

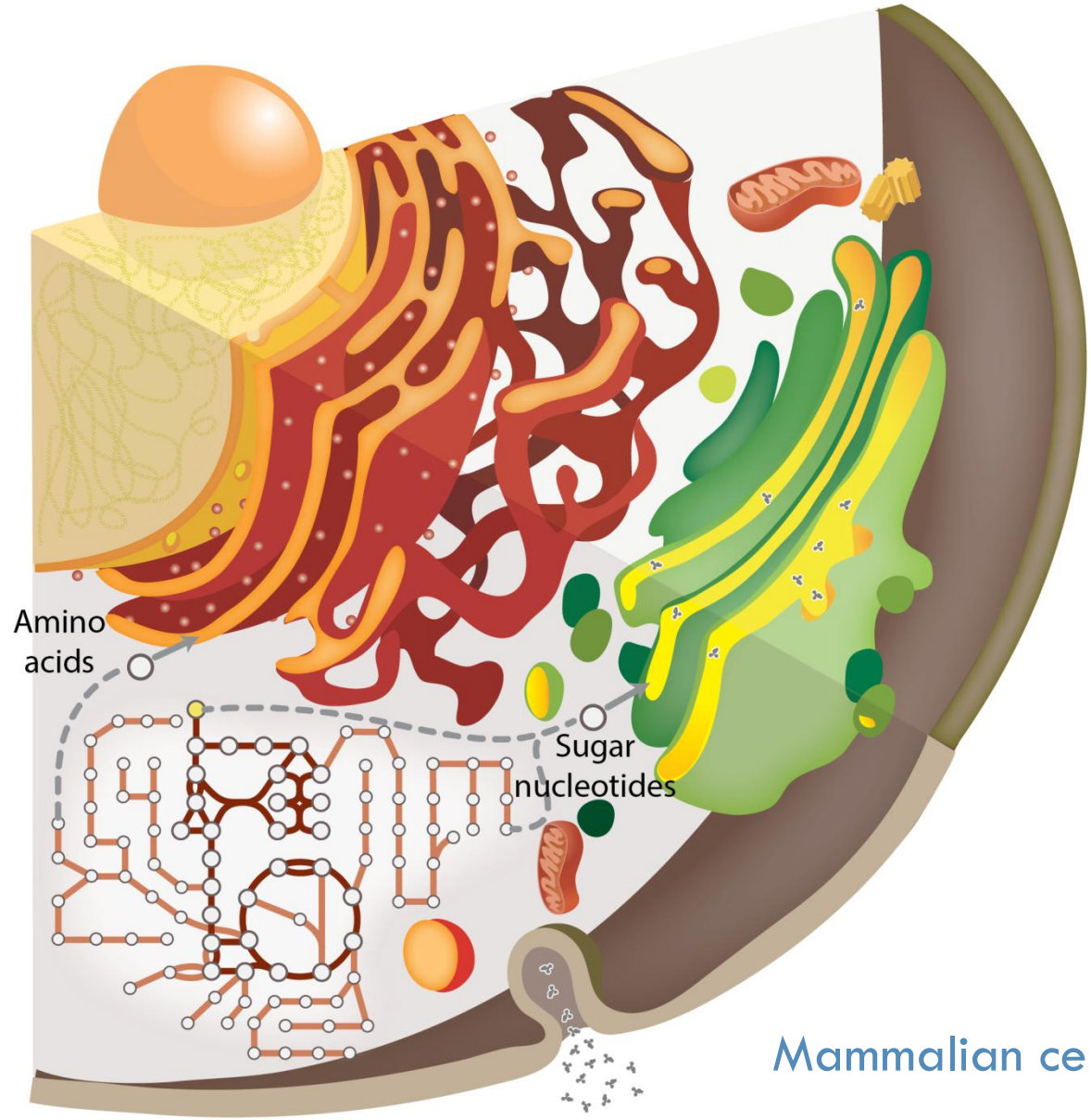


Kumar *et al.*, FEBS Letters (2009)

LABELING FOR QUANTIFICATION

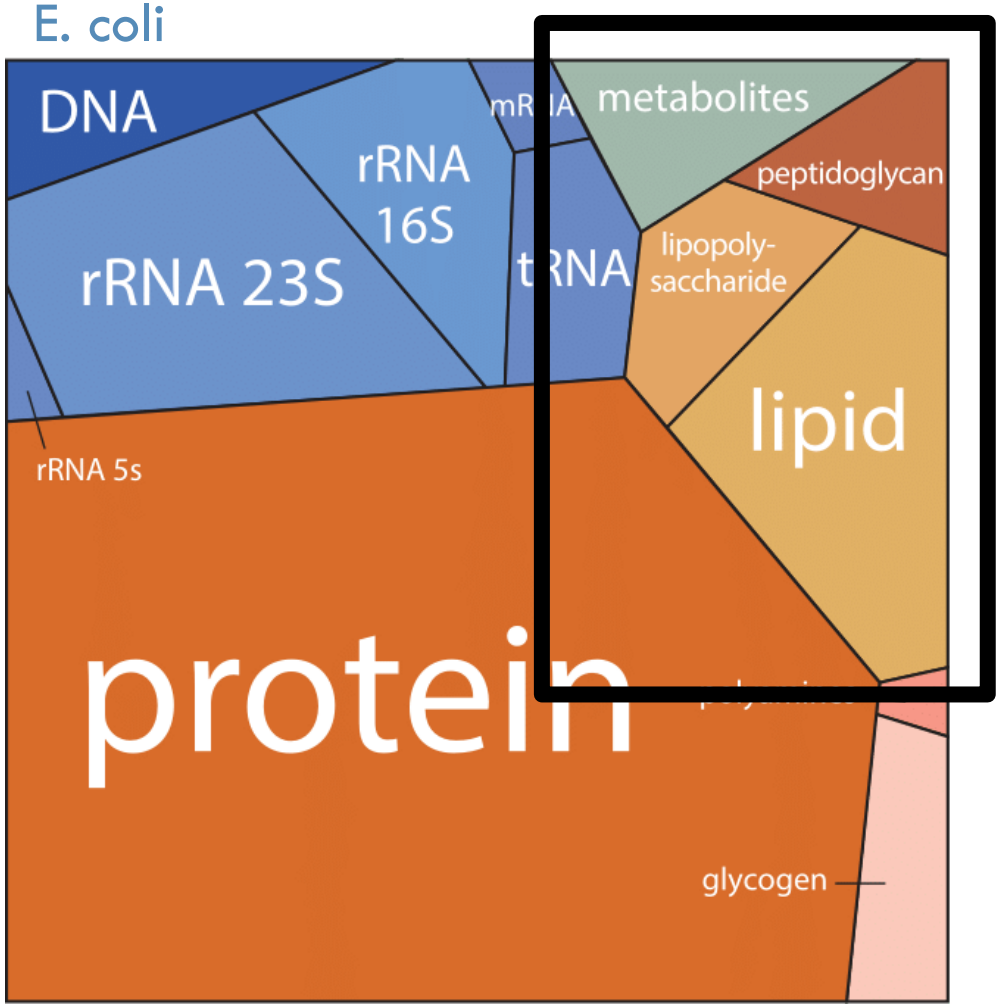


WHAT ARE THE CELL PARTS?



Mammalian cell

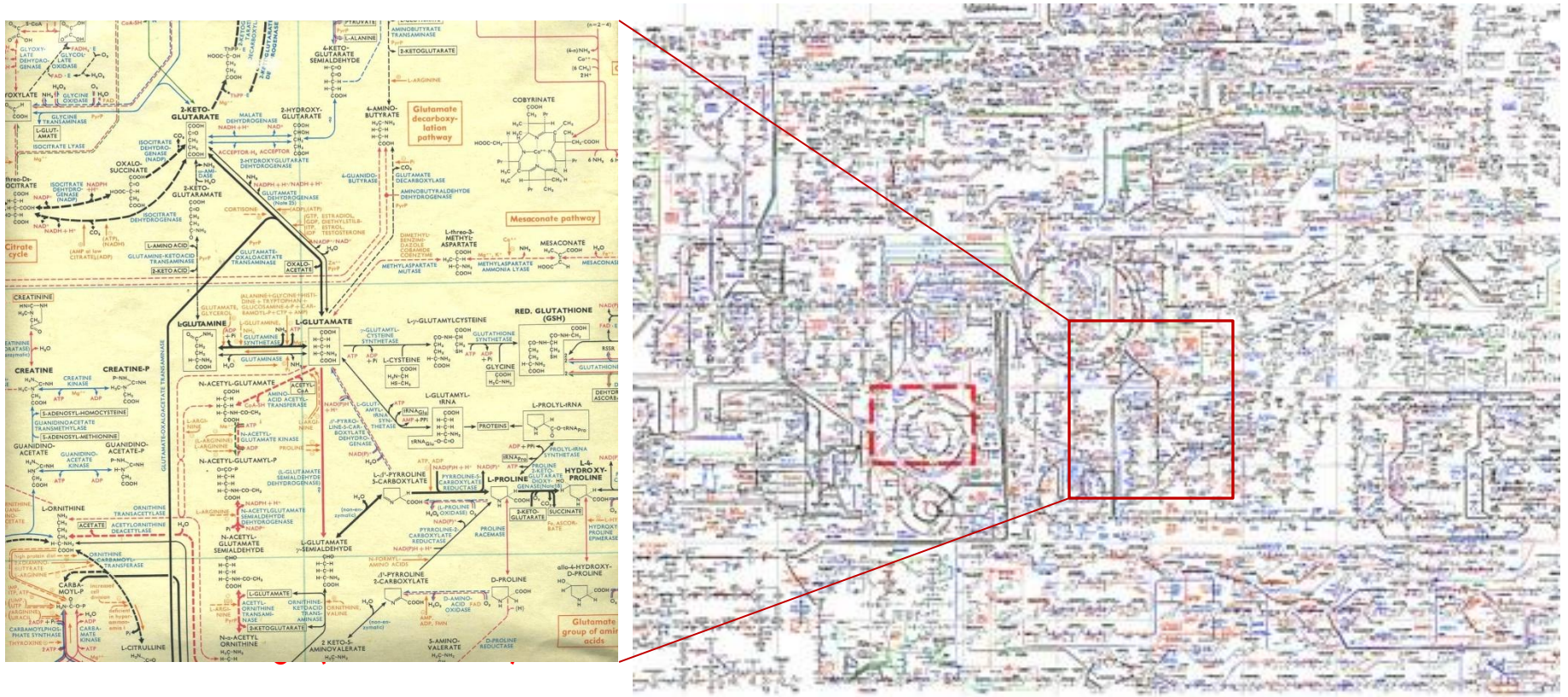
E. coli



Metabolomics

What is metabolism? How much of an organism's genome is dedicated to it? What is a metabolite?

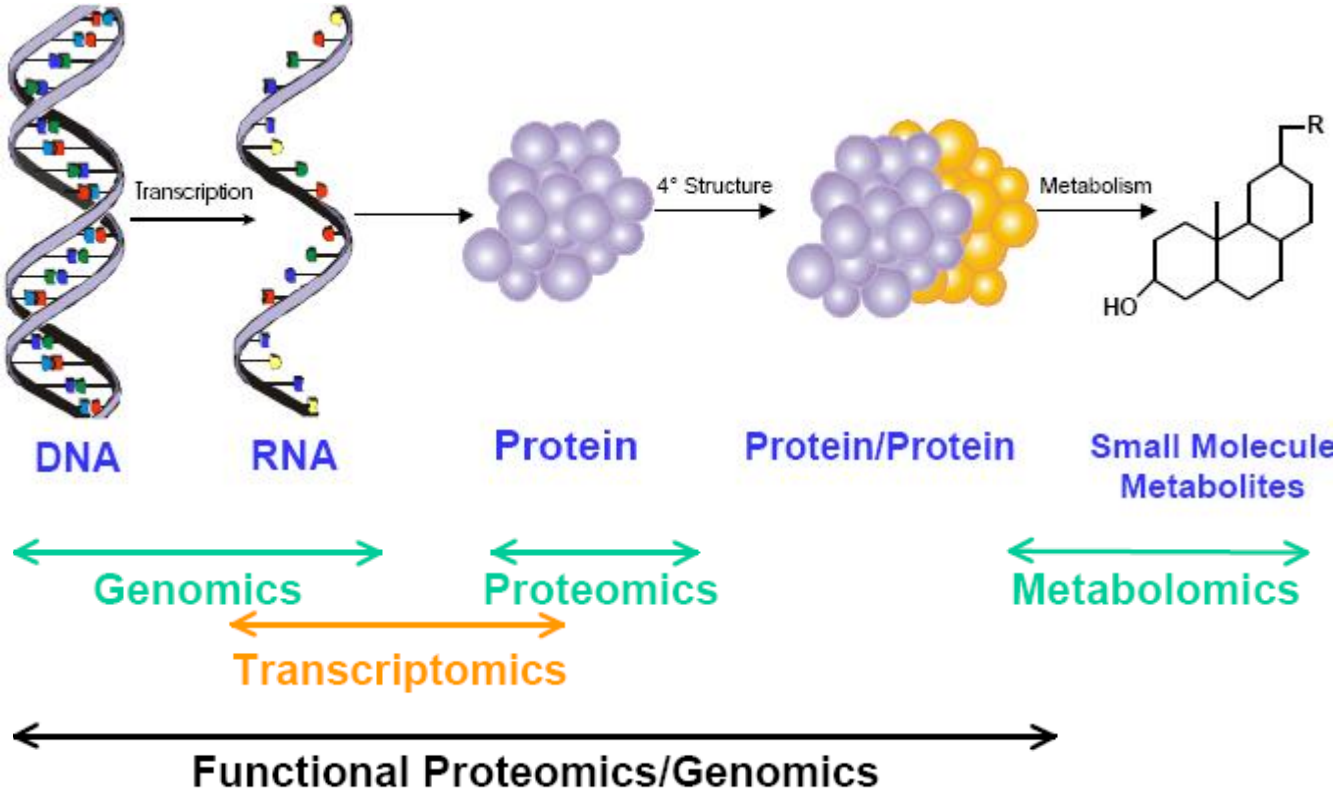
- A low molecular weight (MW < 1000 Da) organic molecule in an organism or biological sample.



Metabolome = the total metabolite pool

Metabolomics = high-throughput analysis of metabolites

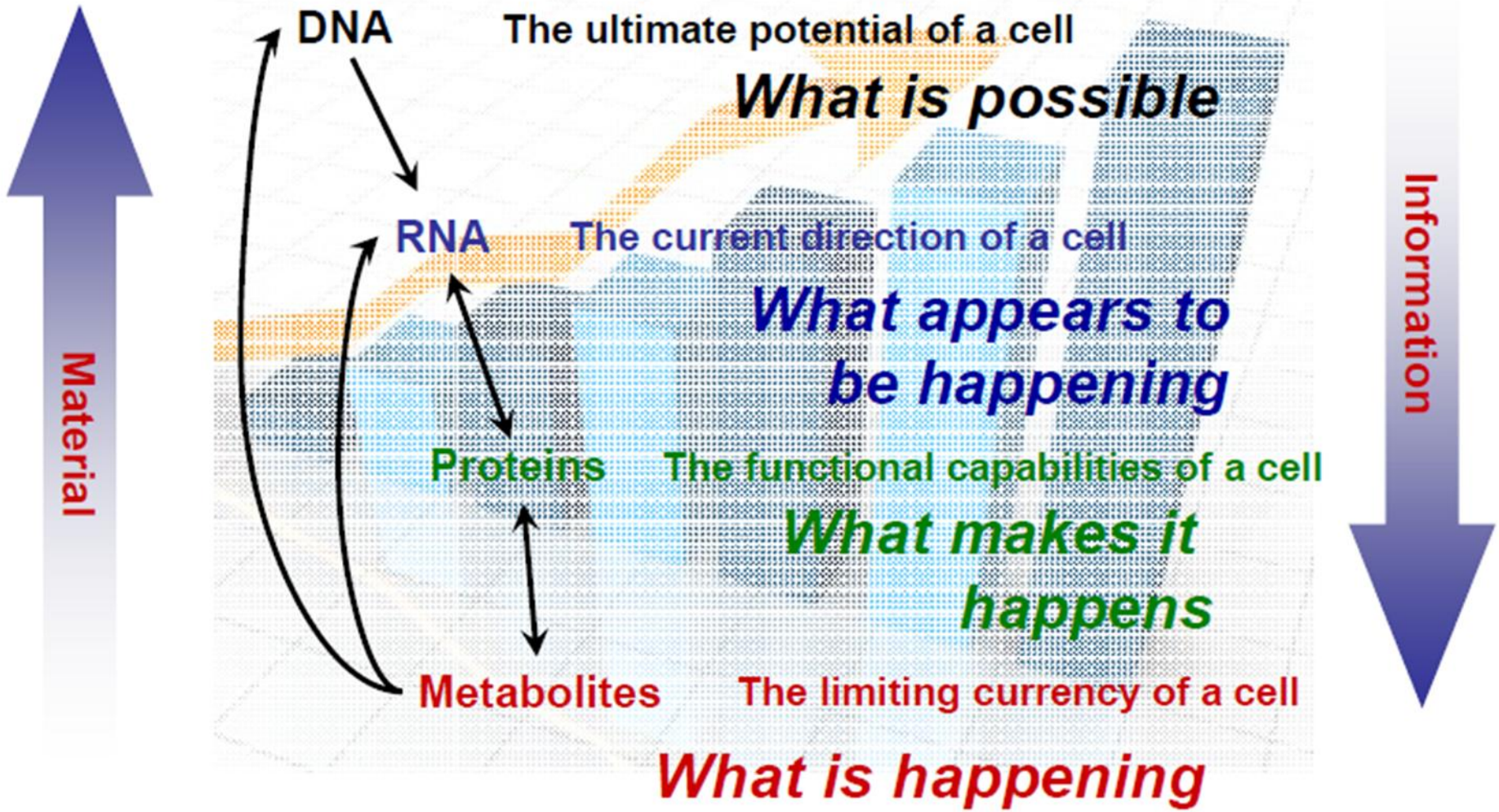
Metabolomics is the simultaneous ('multiparallel') measurement of the levels of many cellular metabolites (hundreds+). Many of these are not identified (i.e. are just peaks in a profile).



Metabolomics = high-throughput analysis of metabolites

Metabolomics analysis is like a snapshot, showing which compounds are present and at what relative levels at a specific time point.

More generally, metabolomics refers to a holistic analytical approach to metabolism that is not guided by specific hypotheses. Instead, metabolomics sets out to determine how (in principle, all) metabolite levels respond to genetic or environmental changes and, from the data, to generate new hypotheses.


DNA
The ultimate potential of a cell
What is possible
RNA
The current direction of a cell
What appears to be happening
Proteins
The functional capabilities of a cell
What makes it happens
Metabolites
The limiting currency of a cell
What is happening
Material
Information

Metabolomics compared to Genomics, Transcriptomics, and Proteomics

Differences between metabolomics and the other high-throughput data types:

(a) Conceptual:

1 GENE → 1 mRNA → 1 Protein → Many Metabolites

(and conversely: Many proteins → 1 Metabolite)

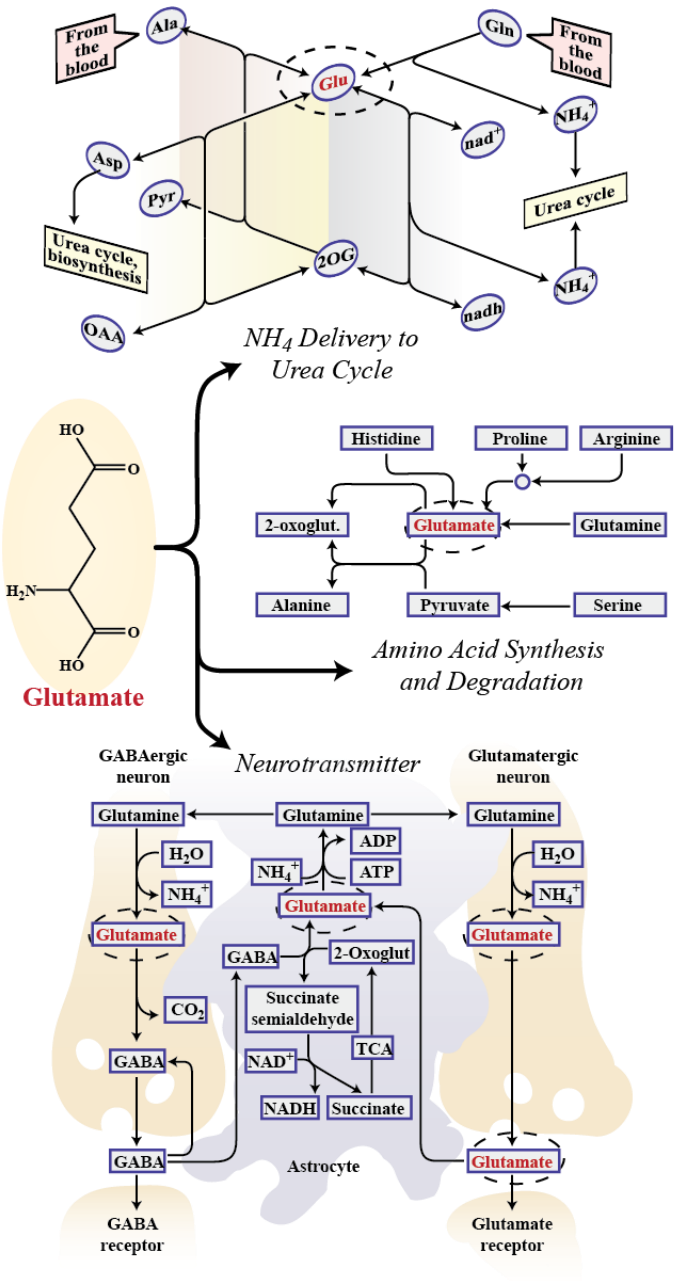
No direct relationship between metabolite and gene (like genes and mRNAs and proteins)

A single gene does not specify the concentration of a single metabolite.

Metabolite concentrations are determined by enzyme activities in pathways.

In practice, therefore, metabolite levels change according to developmental, physiological, and pathological states.

Biological variance in metabolite levels between genetically identical organisms in the same conditions is large – about 10× the analytical variability – and limits the resolution of metabolomics.



Metabolomics compared to Genomics, Transcriptomics, and Proteomics

Differences between metabolomics and the other high throughput data types:

(b) Chemical:

Metabolites have the broadest range of diversity in chemical structures and properties. Their molecular weights span two orders of magnitude (30–3000 Da).

No single extraction or analysis method works for all metabolites.
(Unlike DNA sequencing, microarrays, MS analysis of proteins – all are general methods.)

(c) Dynamic:

Many metabolite levels change with half times of minutes or seconds .

Valuable information is lost if sampling times are too far apart.

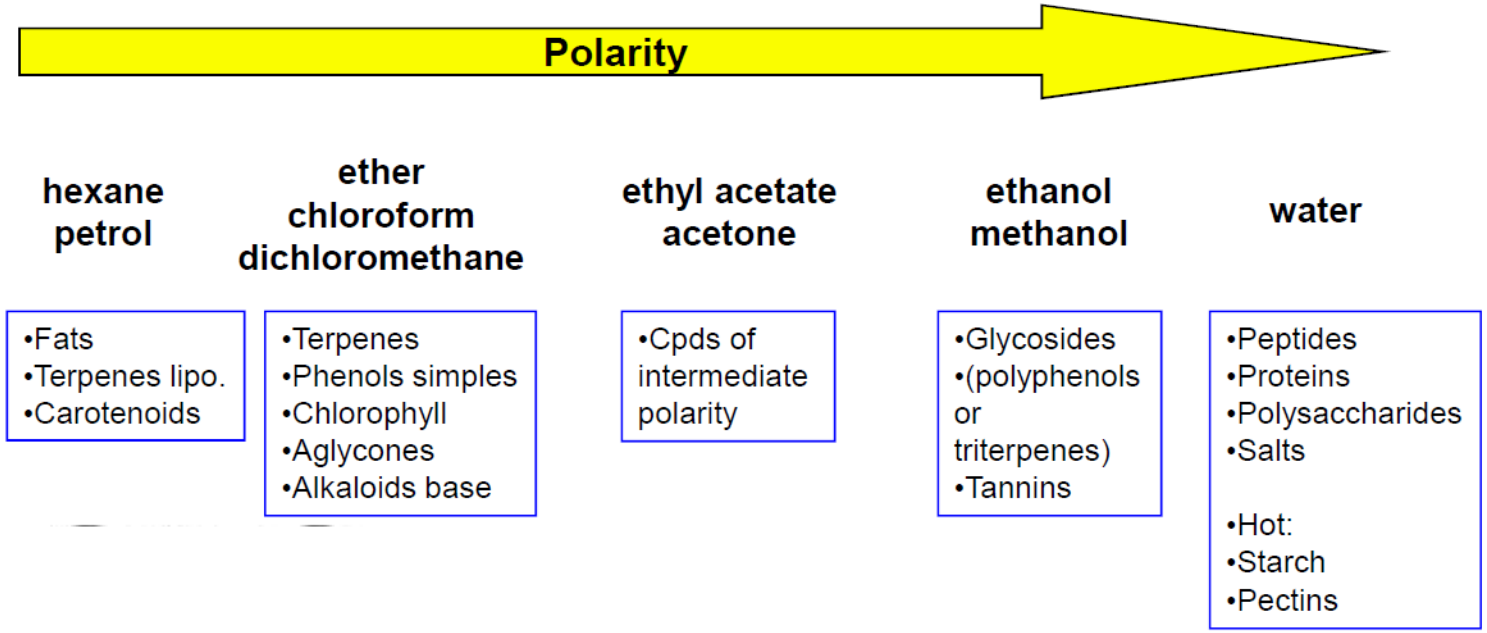
Drastic artifactual changes can occur in short intervals between harvest and extraction; this adds to biological variance.

Metabolic Profiling Methods

Sample Preparation

Metabolites are typically extracted in aqueous or methanolic media, then fractionated into lipophilic and polar phases that are then analyzed separately. Further fractionation of each phase may follow to split metabolites into classes prior to analysis.

No single extraction procedure works for all metabolites because conditions that stabilize one type of compound will destroy other types or interfere with their analysis. Therefore the extraction protocol has to be tailored to the metabolites to be profiled.



Metabolic Profiling Methods

Sample Preparation

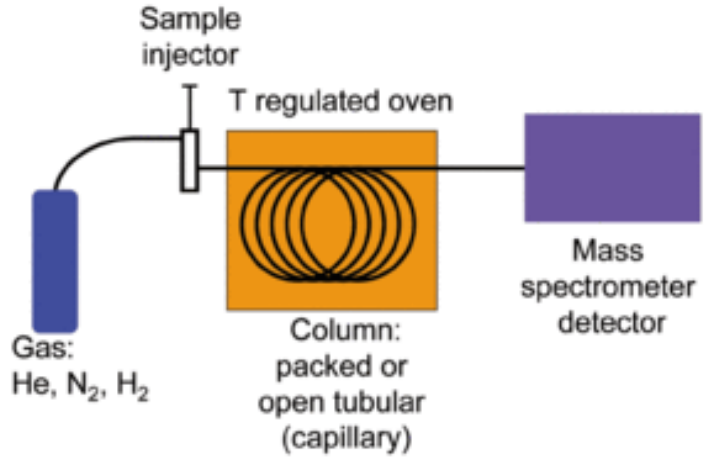
In practice, these considerations mean that metabolic profiling is often confined to fairly stable compounds that can be extracted together. These include major primary metabolites (sugars, sugar phosphates, amino acids, and organic acids) and certain secondary metabolites (e.g., phenylpropanoids, alkaloids).

The most comprehensive profiling can cover several hundred such compounds, many of which are unidentified. Many crucial metabolites, particularly minor or unstable ones, are currently being missed in metabolomics analyses.

Metabolic Profiling Methods

Main Analytical Techniques

Gas Chromatography/Mass-Spectrometry (GC/MS)



In GC/MS, it may be necessary to first derivatize the sample to increase metabolite stability and volatility. The derivatized mix is then fractionated by a gas chromatograph that is coupled to a mass spectrometer.

The mass spectrometer scans the peaks emerging from the GC column at frequent intervals (~1 sec) and so acquires the mass spectrum of each peak, from which peaks can be identified and quantified. Mass spectrometry 'weighs' ionized individual molecules and their fragments. Molecules are identified from their fragmentation pattern and 'weights' (mass/charge ratios – m/z values), with the help of mass spectra libraries, and can be quantified from peak size.

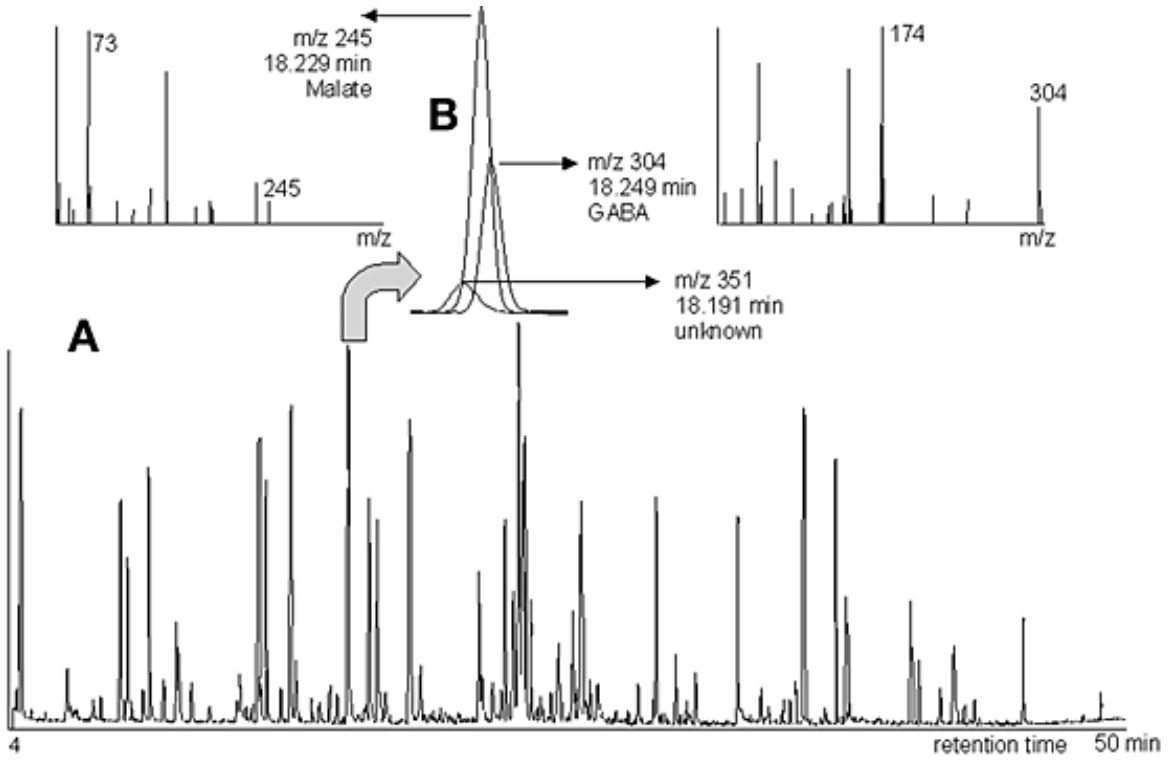
Metabolic Profiling Methods

Main Analytical Techniques

Gas Chromatography/Mass-Spectrometry (GC/MS)

Overlapping peaks can be deconvoluted because the spectra of their constituents are distinct

Target metabolites are identified by exact retention times and their corresponding mass spectra (B) as shown for the co-eluting peaks of malate, gamma-aminobutyric acid (GABA), and an unidentified compound. m/z, Ratio of mass to charge.
PMID: 11062433



Metabolic Profiling Methods

Main Analytical Techniques

Liquid Chromatography/Mass-Spectrometry (LC/MS)

In LC/MS (also termed high performance liquid chromatography, HPLC/MS) the samples are not derivatized before analysis and an HPLC instrument is used for separation. LC/MS is more suitable than GC/MS for labile compounds, for those that are hard to derivatize, or hard to render volatile. LC/MS is less developed than GC/MS. A closely related method is capillary electrophoresis (CE)/MS.

Metabolic Profiling Methods

Main Analytical Techniques

Liquid Chromatography/Mass-Spectrometry (LC/MS)

Profiling example: Metabolites related to plant isoprenoid biosynthesis. The total ion chromatogram (TIC) is the total output of the ion detector; the extracted ion chromatograms (EICs) are the outputs for particular ions characteristic of isoprenoid synthesis intermediates.

LC-MS analysis of endogenous pools of prenyl diphosphates in isolated peppermint oil gland secretory cells.

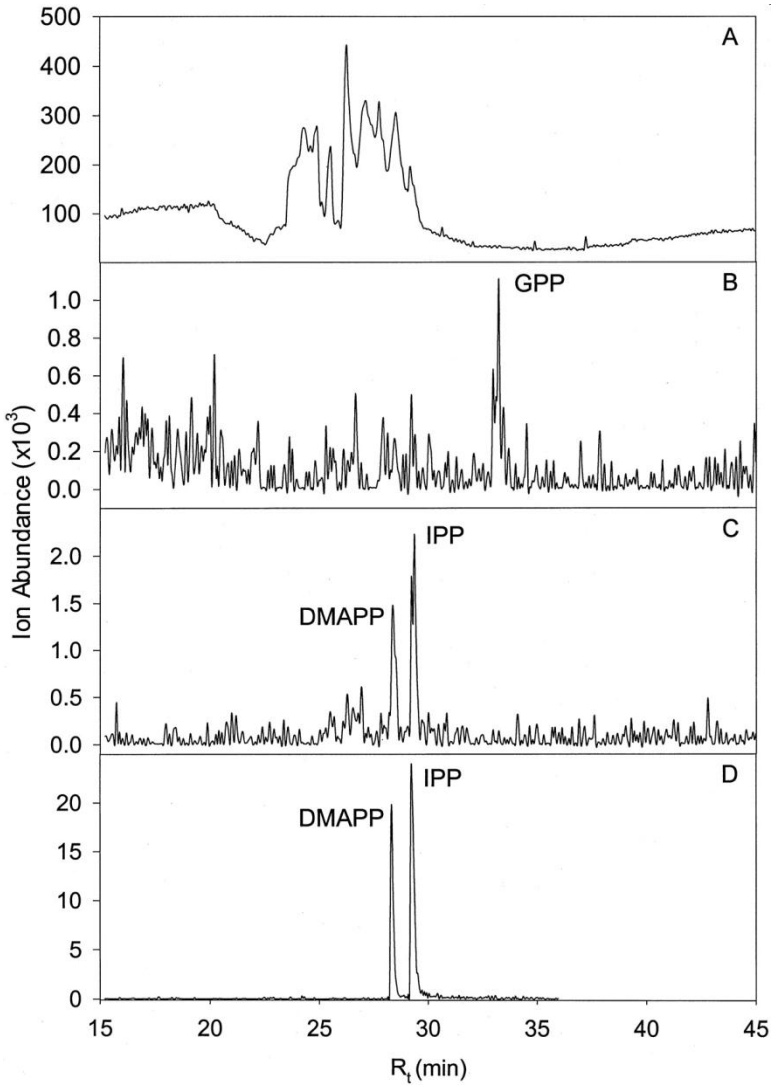
A, Total ion chromatogram (TIC; m/z 50–350)

B, detection of endogenous GPP in the m/z 313 [(M − H)−] extracted ion chromatogram (EIC)

C, detection of endogenous DMAPP and IPP in the m/z 245 [(M − H)−] EIC

D, EIC of a mixture of authentic DMAPP and IPP standards at m/z 245 [(M − H)−].

PMID: 11553758

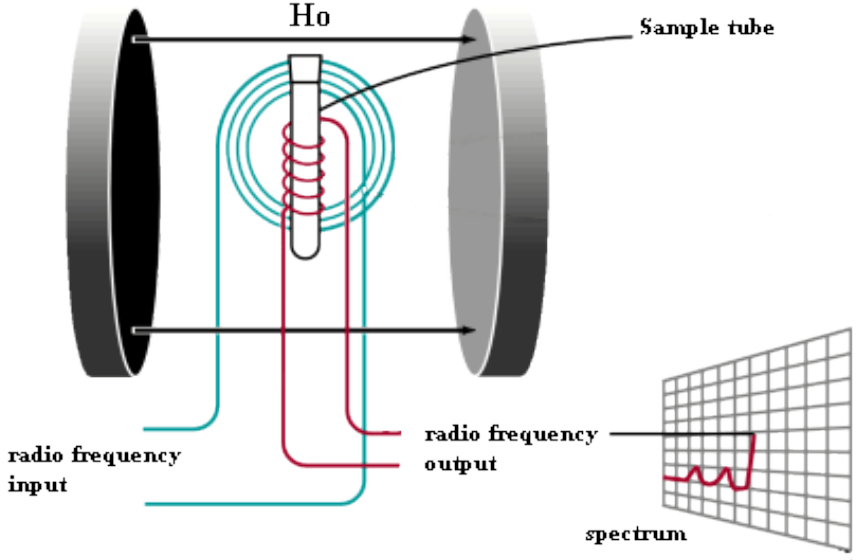


Metabolic Profiling Methods

Main Analytical Techniques

Nuclear Magnetic Resonance (NMR) Spectroscopy

NMR uses radio-frequency (RF) radiation and magnetic fields. RF radiation is used to stimulate nuclei present within molecules. The information obtained is displayed as a spectrum. The horizontal axis is the chemical shift (delta, in units of ppm), which is a measure of the position at which RF absorption occurs relative to an internal standard (tetramethylsilane, TMS). The vertical axis is the intensity of the absorption. As with other spectral techniques, compounds have characteristic spectra. More than 100 metabolites occur in plants at levels high enough for analysis by NMR, so NMR spectra of mixtures contain many peaks.



Metabolic Profiling Methods

Main Analytical Techniques

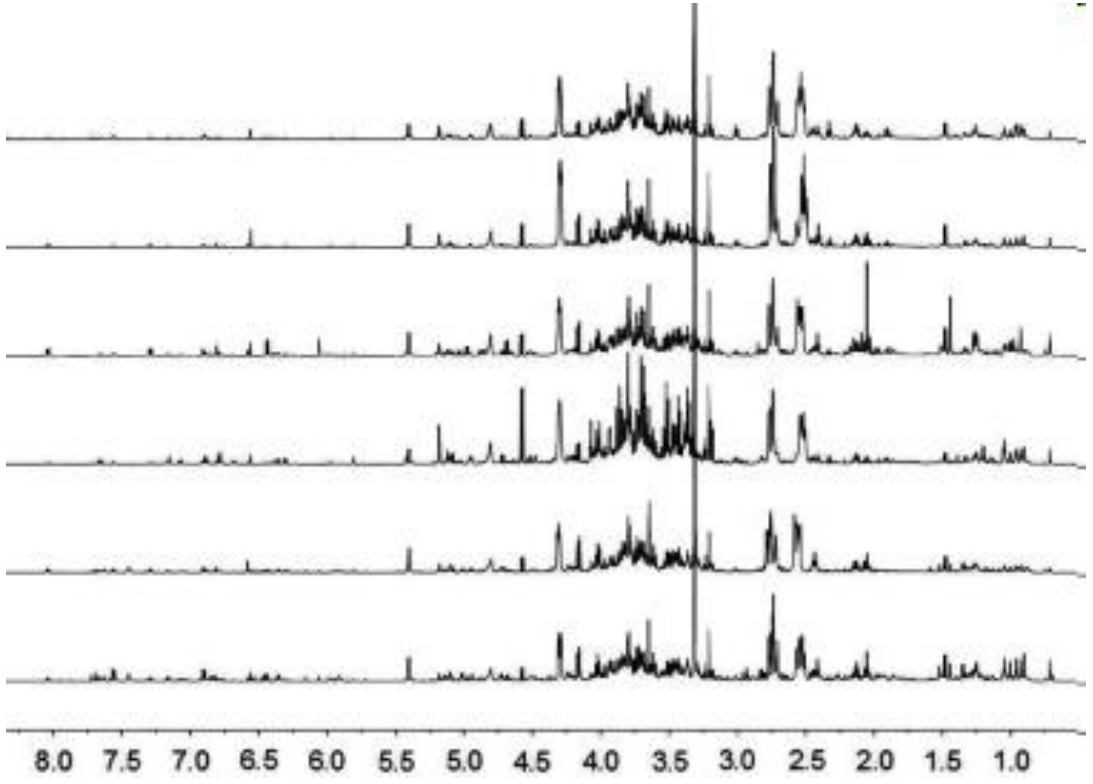
Nuclear Magnetic Resonance (NMR) Spectroscopy

Profiling example: ^{1}H-NMR spectra of extracts of leaves of various *Verbascum* species (medicinal plants)

600 MHz ^{1}H NMR spectra of extracts of *Verbascum* leaves.

From bottom to top:
V. xanthophoeniceum, *V. nigrum*, *V. phlomoides*, *V. phoeniceum*, *V. phlomoides*, *V. densiflorum*.

PMID: 21807390

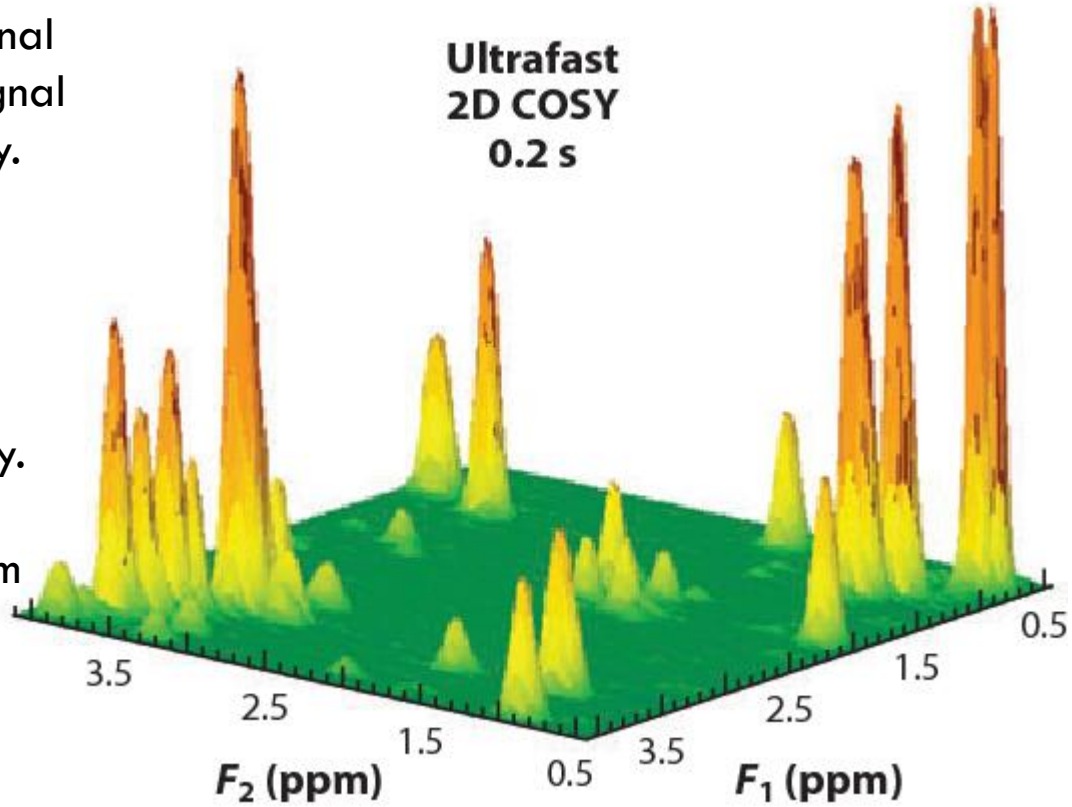


Metabolic Profiling Methods

Main Analytical Techniques

Nuclear Magnetic Resonance (NMR) Spectroscopy

Signal overlap is a problem in the complex spectra of cell extracts. Signal overlap hampers metabolite identification and quantification. Better signal resolution can be obtained using various types of 2D NMR spectroscopy. These approaches cut signal overlap by spreading the resonances in a second dimension.



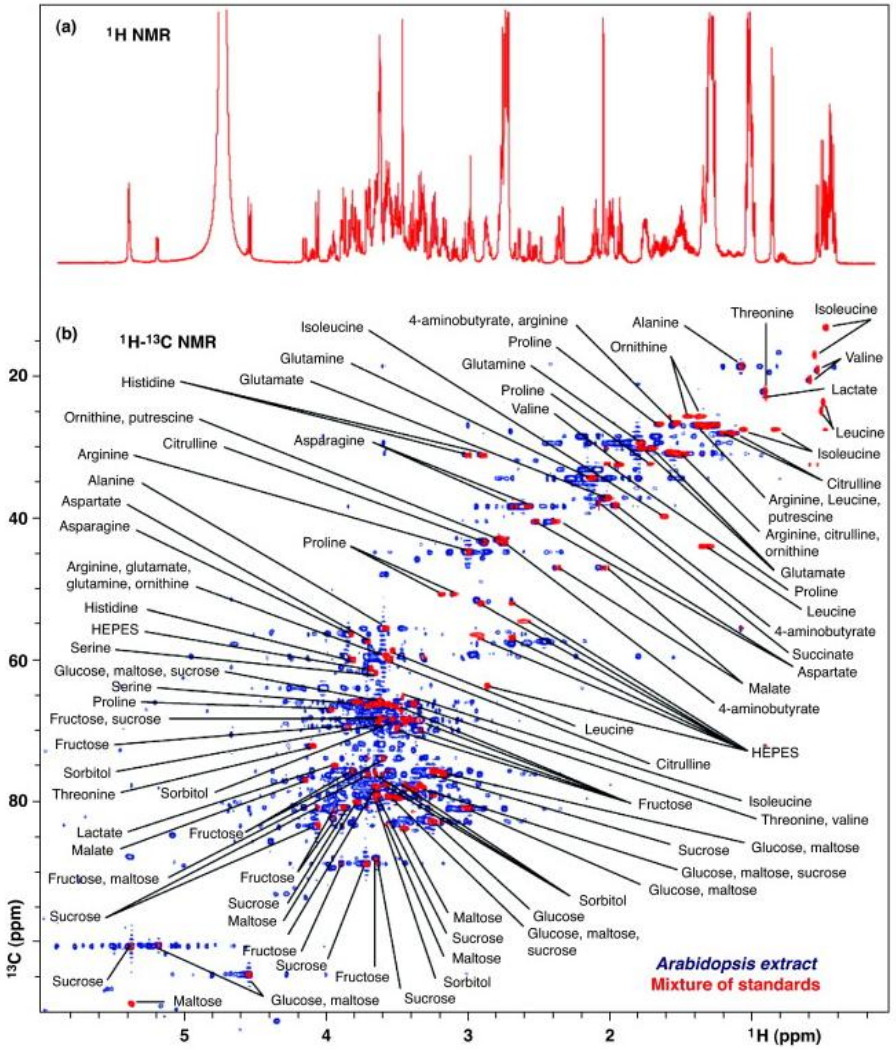
Example: Heteronuclear single quantum coherence (HSQC) spectroscopy. The 2D spectrum has one axis for ^{1}H and the other for a *heteronucleus* (an atomic nucleus other than a proton), usually ^{13}C or ^{15}N. The spectrum contains a peak for each unique proton attached to the heteronucleus being considered.

NMR tutorial: http://www.cis.rit.edu/htbooks/nmr/

Metabolic Profiling Methods

Main Analytical Techniques

Nuclear Magnetic Resonance (NMR) Spectroscopy



HSQC used to select for protons directly bonded to ^{13}C.

Use of HSQC spectroscopy for analysis of common metabolites. In 1D spectra, overlapped signals hamper identification of individual metabolites, whereas in 2D correlation, spots are easily visible.

(**a**) 1D ^{1}H NMR spectrum of an equimolar mixture of the 26 standards.

(**b**) 2D ^{1}H–^{13}C HSQC NMR spectra of the same synthetic mixture (red) overlaid onto a spectrum of aqueous whole-plant extract from *Arabidopsis* (blue).

PMID: 21435731

Metabolic Profiling Methods

Main Analytical Techniques

Nuclear Magnetic Resonance (NMR) Spectroscopy

Advantages of NMR over MS:

- NMR does not destroy the sample
- NMR can detect and quantify metabolite because the signal intensity is only determined by the molar concentration
- NMR can provide comprehensive structural information, including stereochemistry

Many atoms have nuclei that are NMR active, but most NMR data are collected for ^{1}H and ^{13}C since these are present in all organic molecules.

The main weakness of NMR is low sensitivity relative to MS. It is therefore less suited for analysis of trace compounds. As the natural abundance of ^{13}C is only 1.1%, ^{13}C-NMR is less sensitive than ^{1}H-NMR. Recent developments have considerably increased sensitivity, making it less of a problem.

Metabolic Profiling Methods

Main Analytical Techniques

How can one decide which analytical platform should be used?

- Should be rapid, reproducible, with easy sample preparation.
- Selection based on objectives, target metabolites, availability, etc.

	HPLC or TLC-UV	GC-MS	LC-MS	MS^n	NMR
Sample preparation	++	−	−	+	+++
Reproducibility	−	+	−	+	+++
Absolute quantitation	−	−	−	−	+++
Relative quantitation	+	++	+	++	+++
Identity	+	++	++	++	++
Compound number	ca. 30	ca. 1000	ca. 200	ca. 1000	ca. 200
Sensitivity	+	++	++	+++	−

Scale from - to +++ for major disadvantages to major advantages

Phytochem Rev (2008) 7:525–537

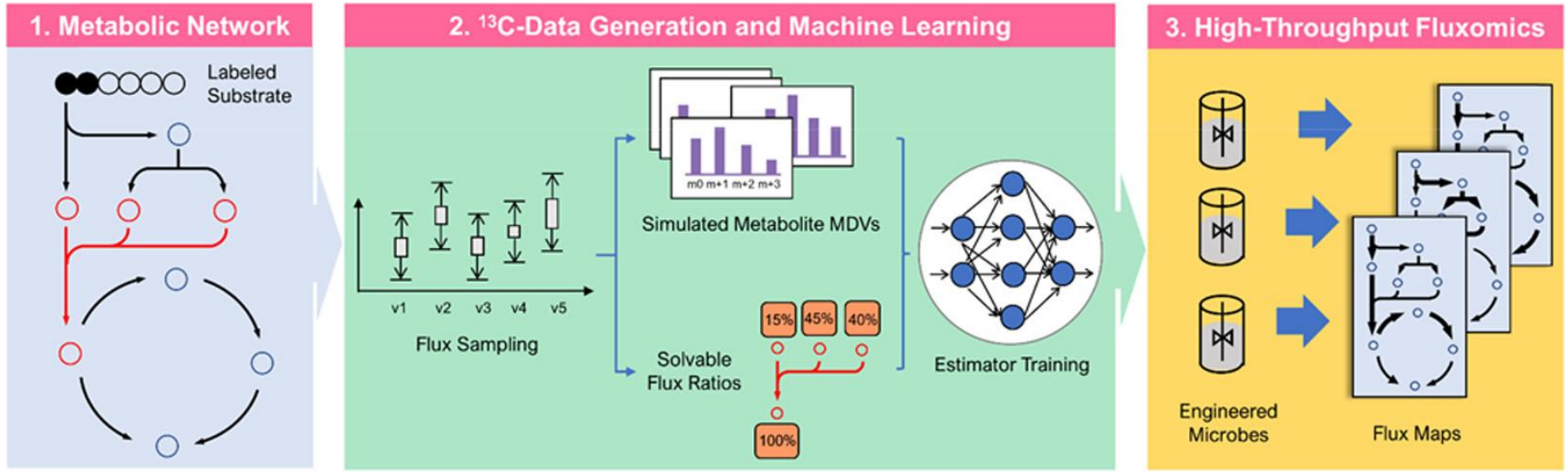
Metabolomics + time = fluxomics?

Fluxomics = A branch of metabolomics that measures the turnover of metabolites in pathways using labeled isotopes such as ^{13}C.

- Concentrations of radio-labeled metabolites are measured at time points

- Using known pathway topology, concentrations are used to infer flow through pathways

- Instead of being a snapshot of metabolism, it is like a movie



Wu, ACS Syn Bio, 2022

What will you do with metabolomics?

- Disease phenotypes
- Drug metabolism and toxicology
- Bioprocess optimization
- Synthetic biology and metabolic engineering

ACTIVITY: DNA, RNA, PROTEINS, OR METABOLITES? WHICH IS OMIC COOLER?

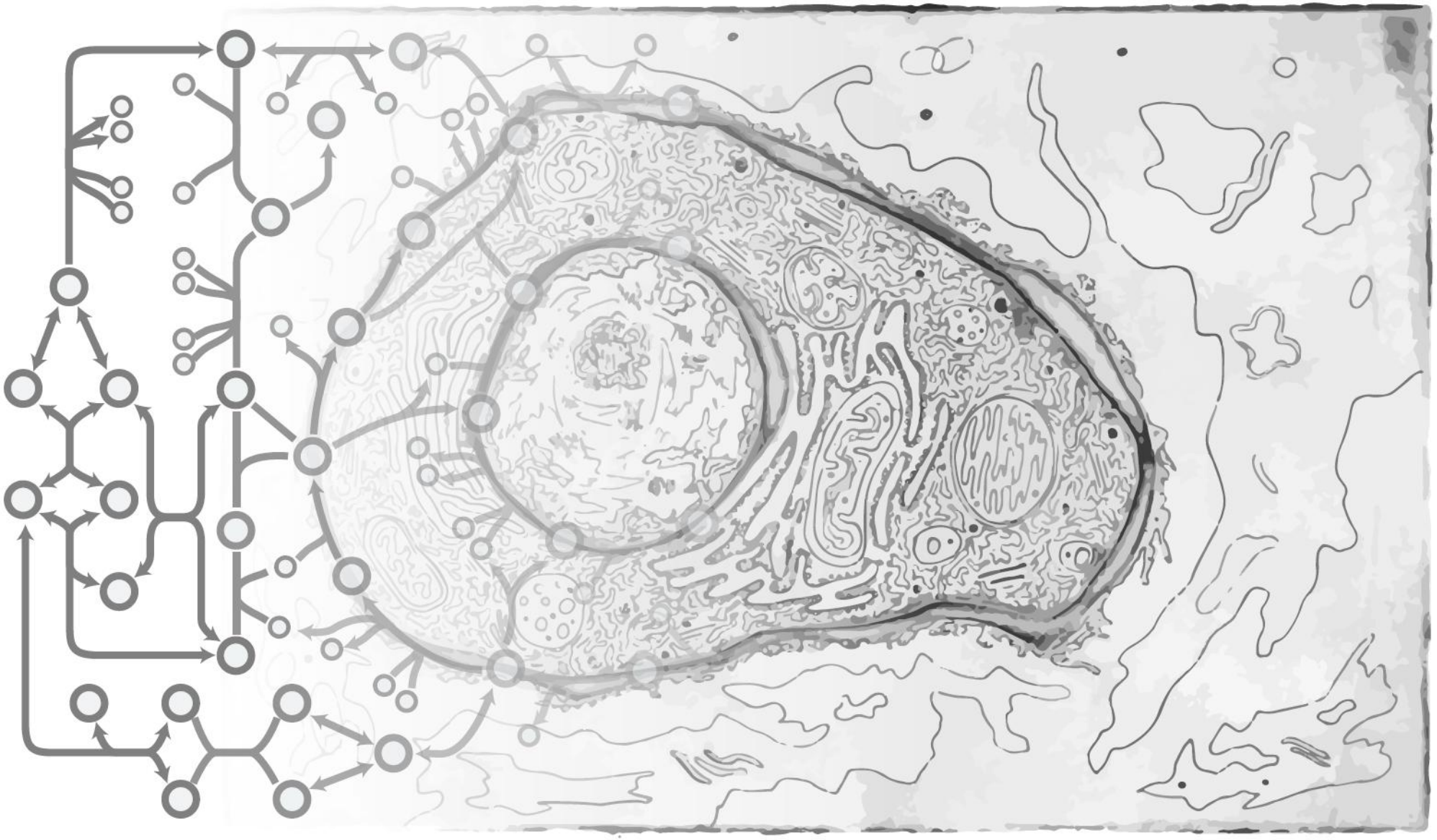
So far we've discussed:

- Genomics
- Transcriptomics
- Proteomics
- Metabolomics

Form groups of 2-3 students

Argue which is the coolest, and why the others are lame.

Give 1 statement backing up your case for each, and ask your favorite chatbot if you're right.



MOLECULAR INTERACTIONS IN THE CELL

CLASSES OF BIOLOGICAL MEASUREMENTS

1) Components

- DNA sequence / genotype:
  Next-gen sequencing, SNP & CNV arrays
- Gene expression:
  DNA microarrays, mRNA sequencing
- Protein levels, locations, mods:
  Mass spectrometry, fluorescence microscopy, protein arrays

2) Interactions

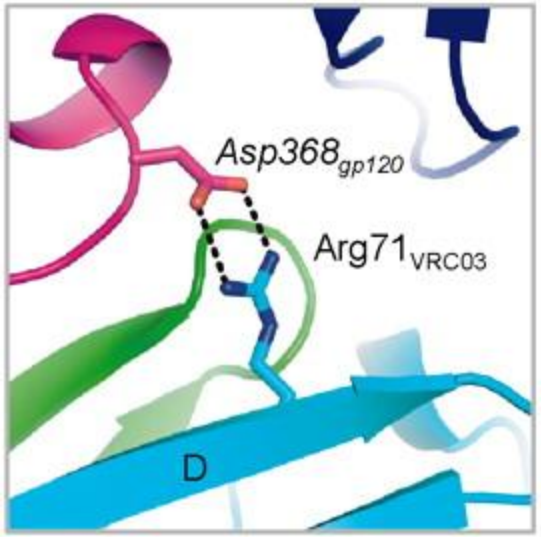
- Protein-protein interactions:
  Two-hybrid system, coIP, protein antibody array
- Protein-DNA interactions:
  Chromatin IP (chip) sequencing
- Protein-compound

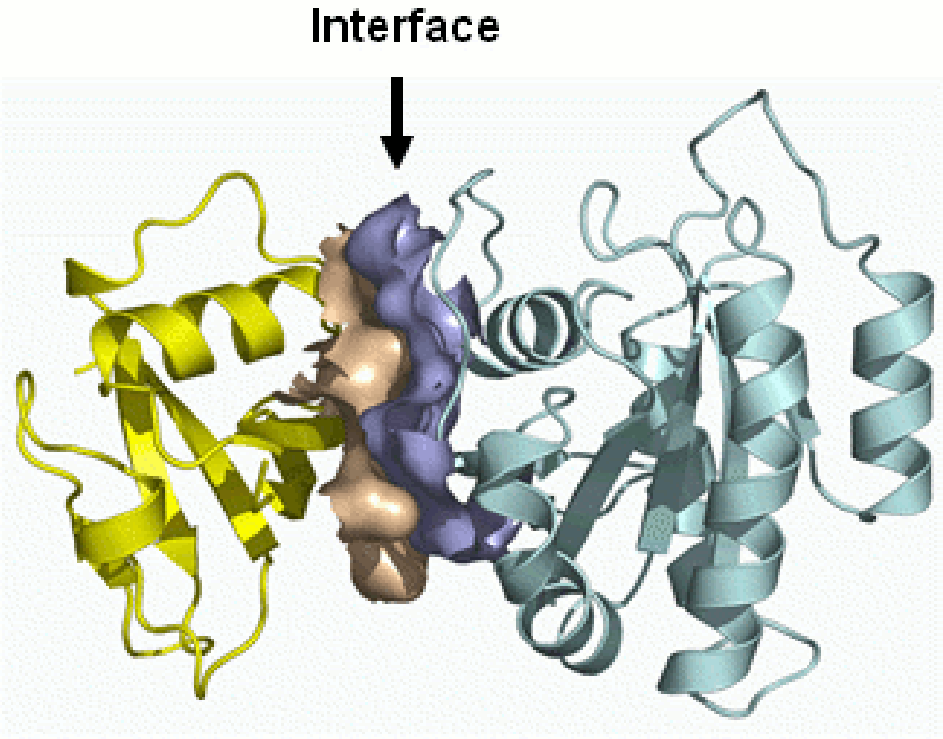
3) Phenotypic traits

- Physiological or disease state, binary or quantitative
- Growth rate, response to stimulus or stress
- Behaviors

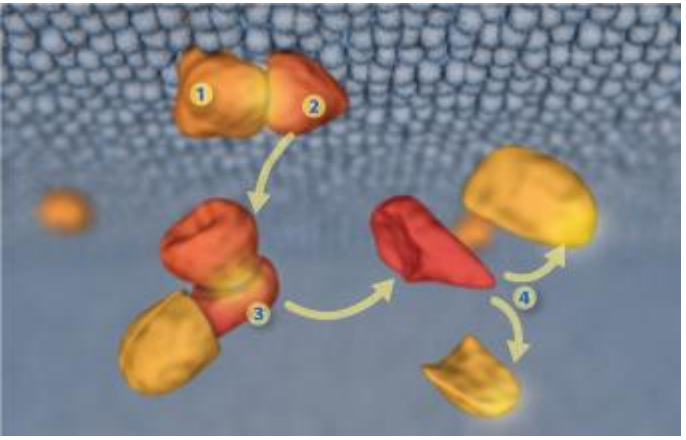
HOW DO PROTEINS INTERACT?

Electrostatic interactions





WHY DO THEY INTERACT?

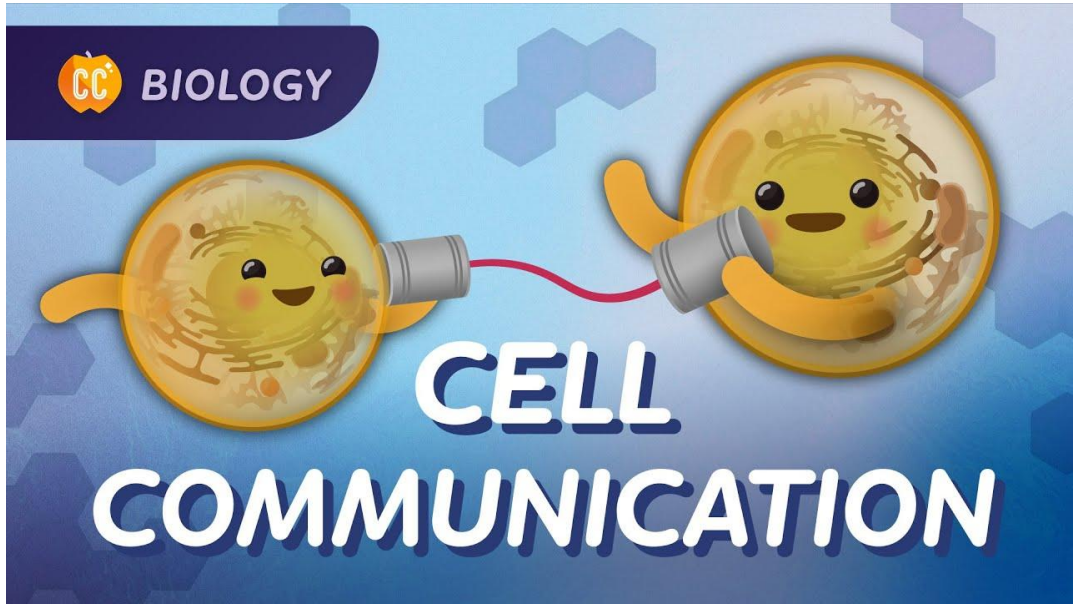


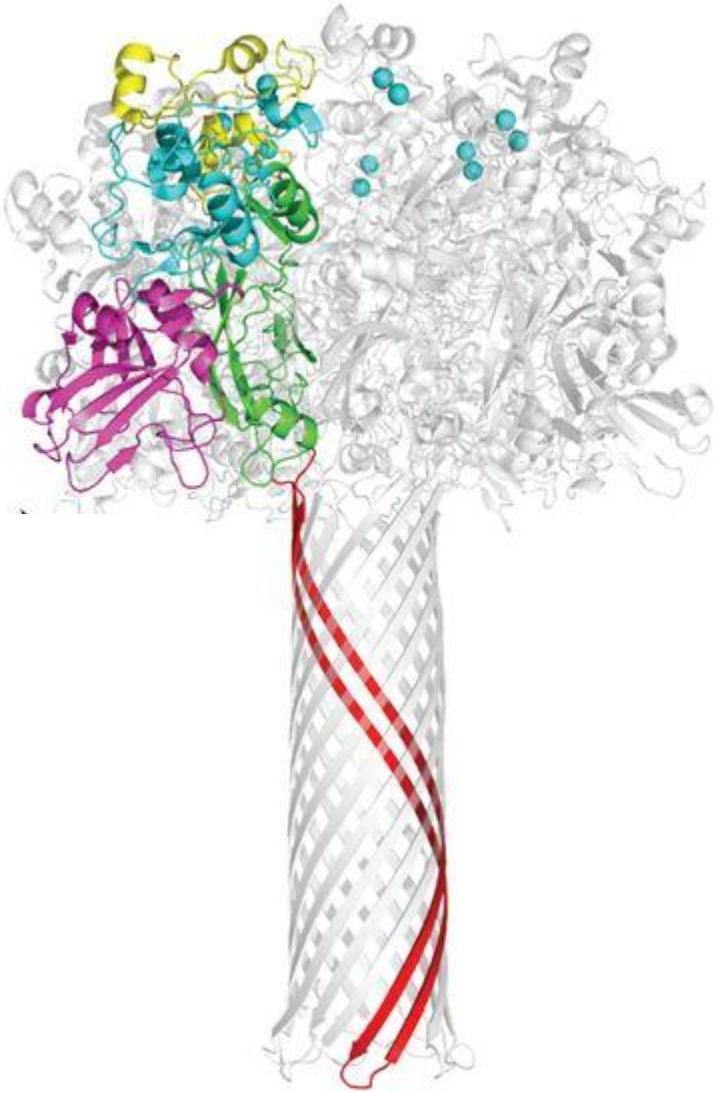
Complexes to work together

Relay and amplify signals

Change protein activity or function

Cell-cell communication



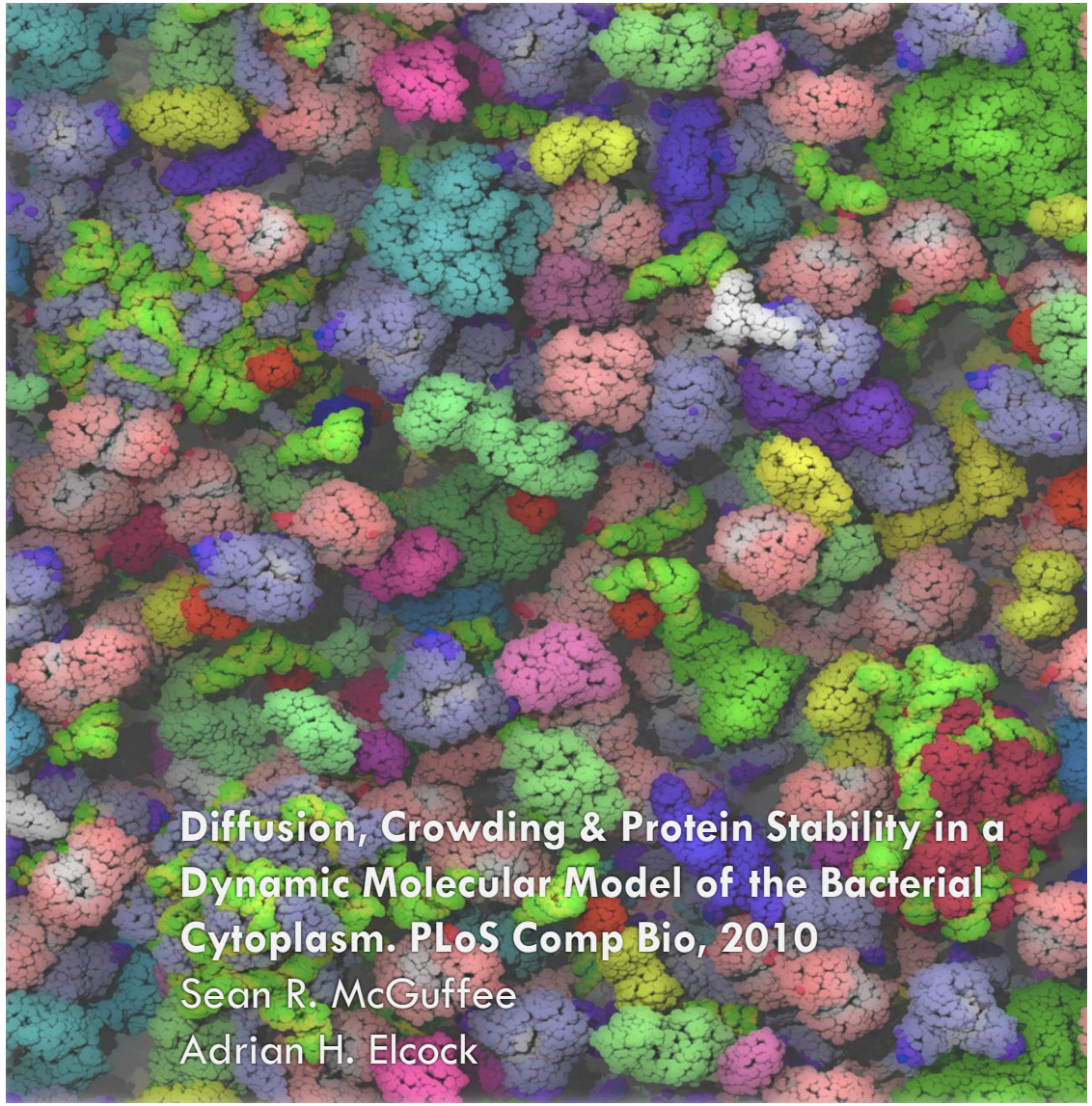


WHY DO THEY NOT INTERACT?

The cell is crowded

- Protein
- RNA
- DNA
- Metabolites

Evolved away negative interactions



https://www.youtube.com/watch?v=zme5VcSYpCg

HIGH-THROUGHPUT METHODS TO DETERMINE PPIS

Yeast 2 hybrid

Phage display

Protein complementation

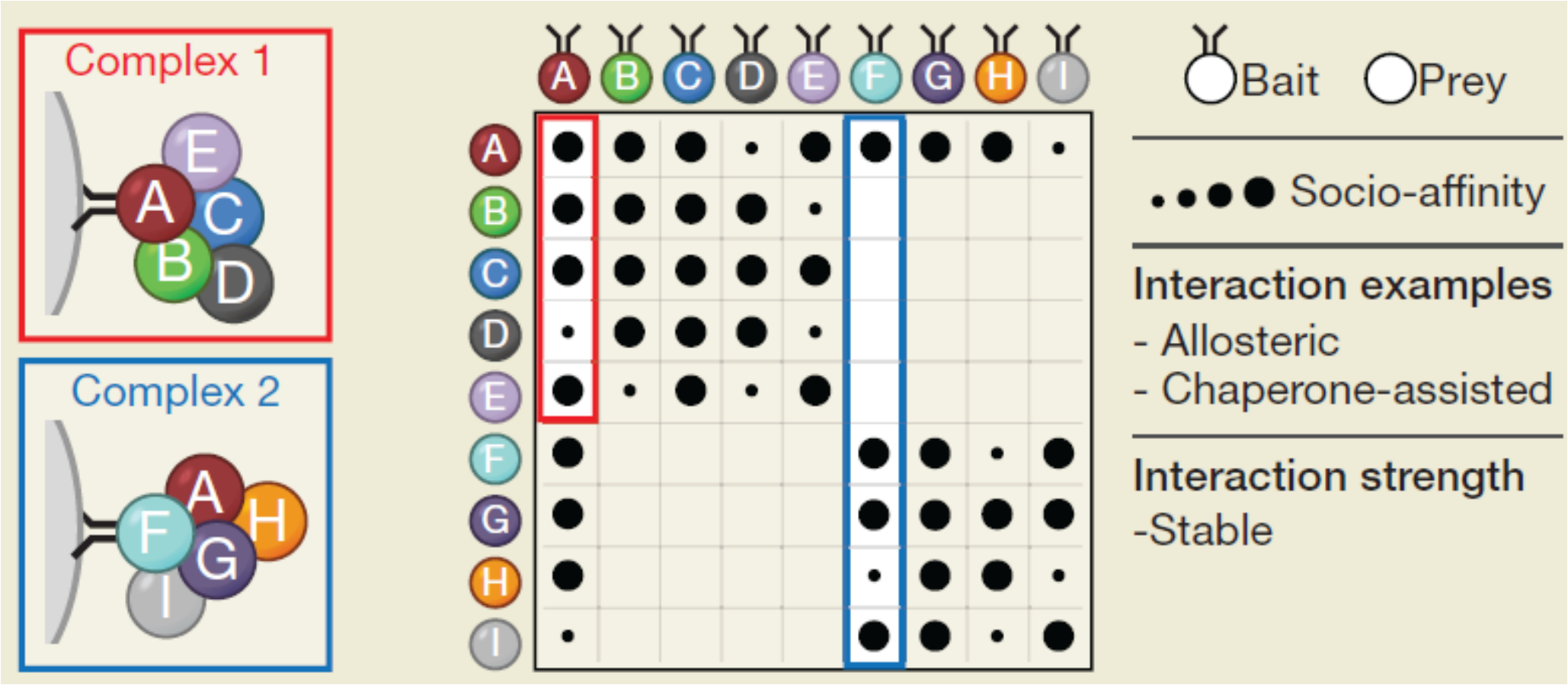
Co-immunoprecipitation

Chemical cross linking

Proximity biotinylation assays

See many more computational and lower throughput methods at “Methods to investigate protein–protein interactions” at wikipedia

Co-IP


Complex 1
E
A
C
B
D
Complex 2
A
H
F
G
I
A B C D E F G H I
Bait
Prey
Socio-affinity
Interaction examples
- Allosteric
- Chaperone-assisted
Interaction strength
-Stable

MIXING THE VARIOUS METHODS… STRING: SEARCH TOOL FOR THE RETRIEVAL OF INTERACTING GENES/PROTEINS

A database of known and predicted protein interactions

Direct (physical) and indirect (functional) associations

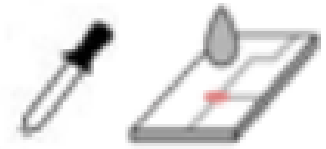
The database currently covers ~~2,483,276~~ 59,309,604 proteins from ~~630~~ 12,535 organisms

The STRING database currently covers proteins from organisms.

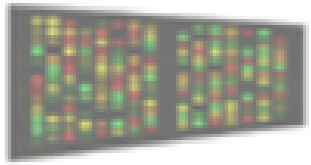
Derived from these sources:

Genomic Context

High-throughput Experiments



(Conserved) Coexpression



Previous Knowledge

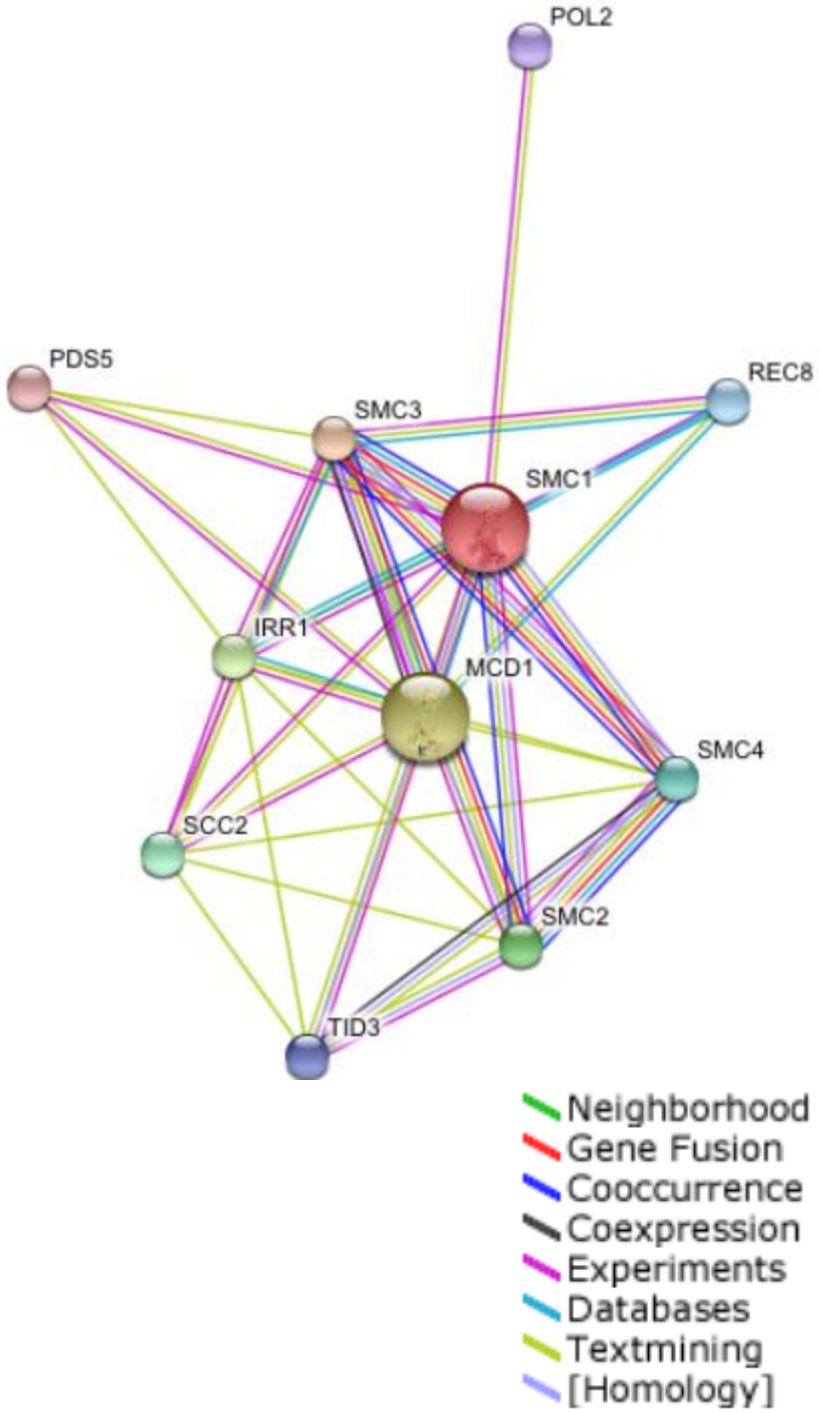
GRAPH OF PPIS

Nodes are proteins

Lines with color is an evidence of interaction between two proteins. The color encodes the method used to detect the interaction.

Click on each node to get the information of the corresponding protein.

Click on each edge to get information of the interaction between two proteins.



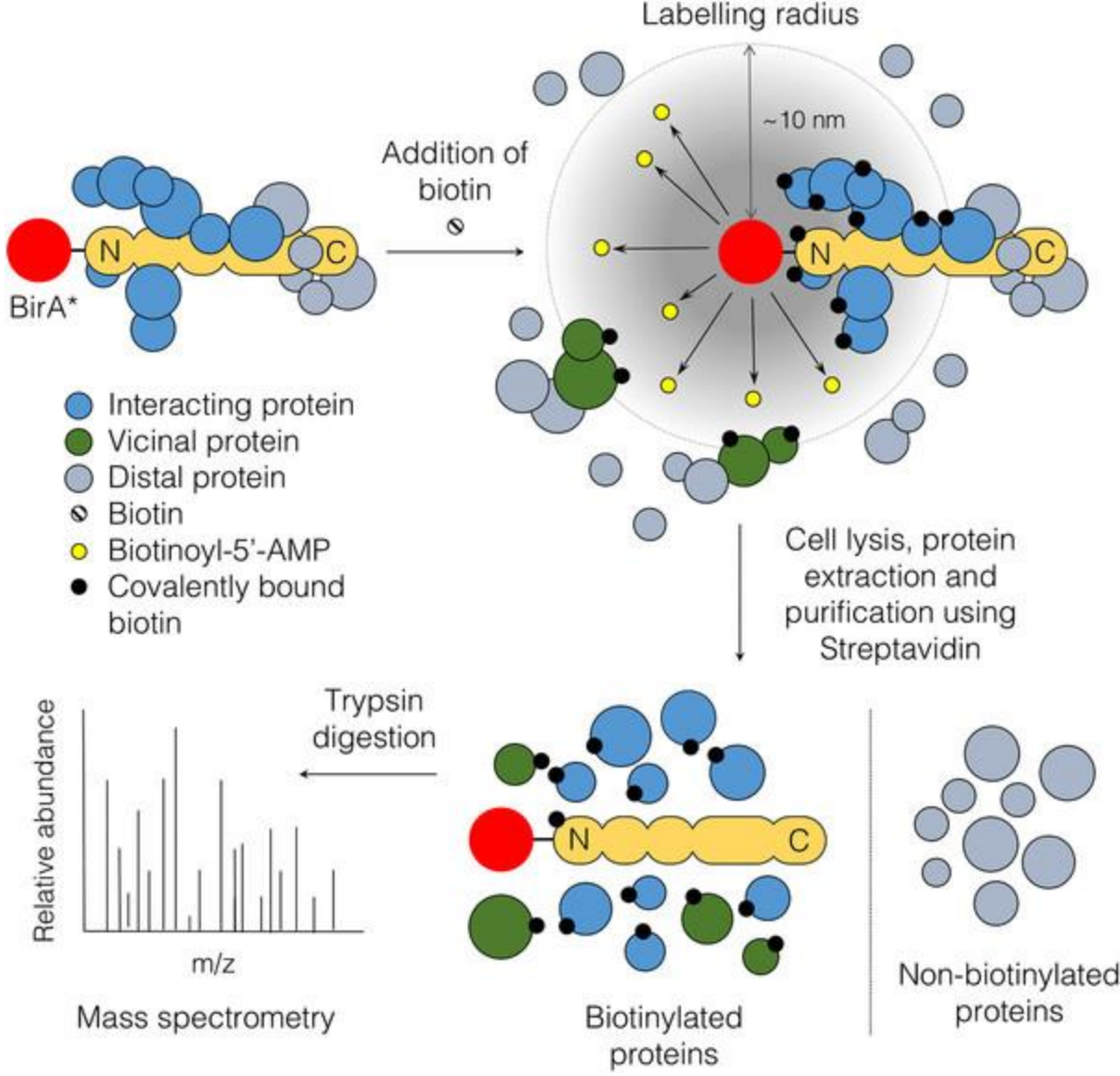
PROXIMITY LABELING

What if we could do everything inside the cell?

Capture transient/weak interactions

Ongoing work to improve size/specificity/etc.

- TurboID
- miniTurbo
- ultraID

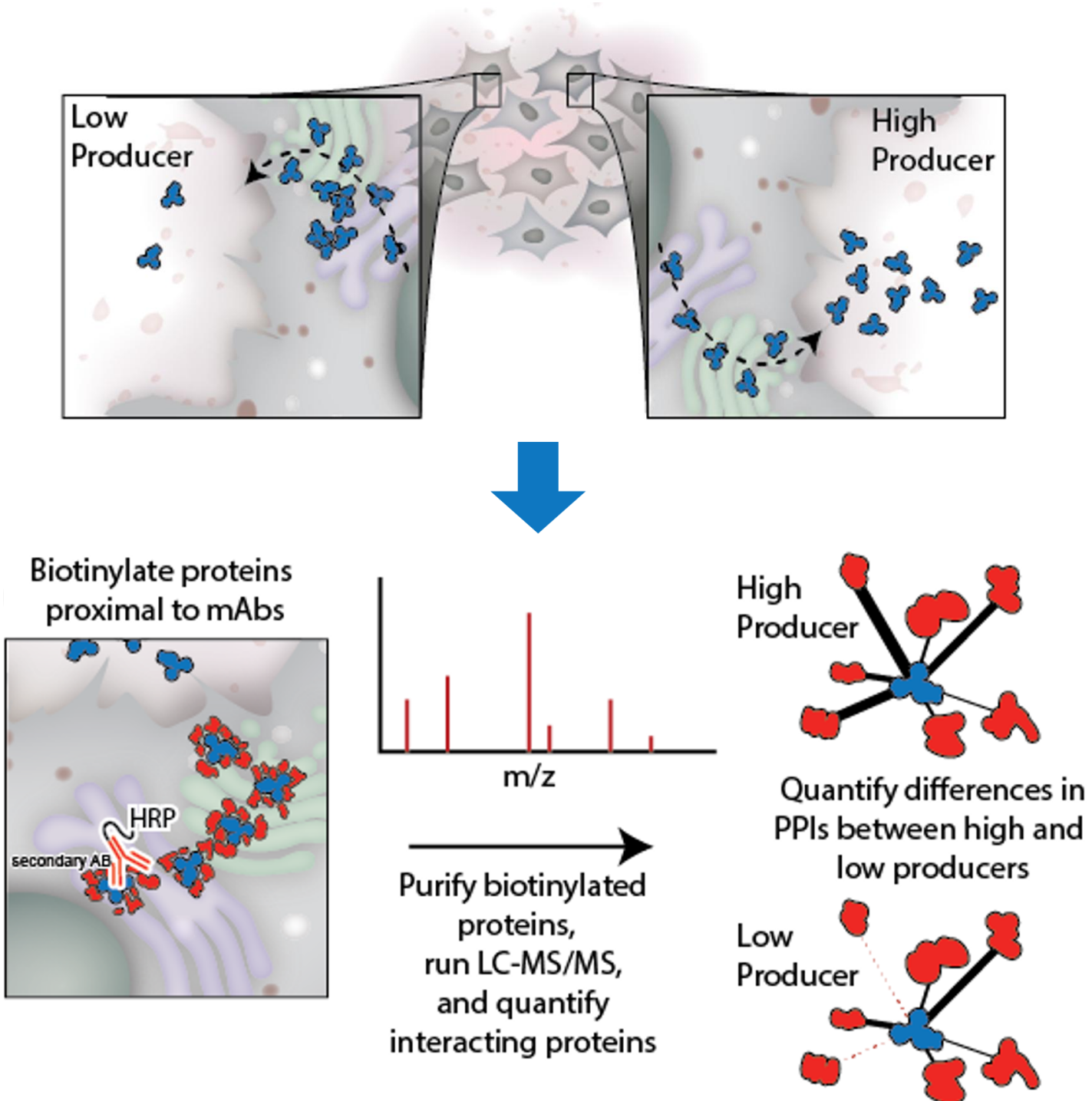


Review > Proteomics. 2016 Oct;16(19):2503-2518. doi: 10.1002/pmic.201600123. Epub 2016 Jul 27.

Meet the neighbors: Mapping local protein interactomes by proximity-dependent labeling with BioID

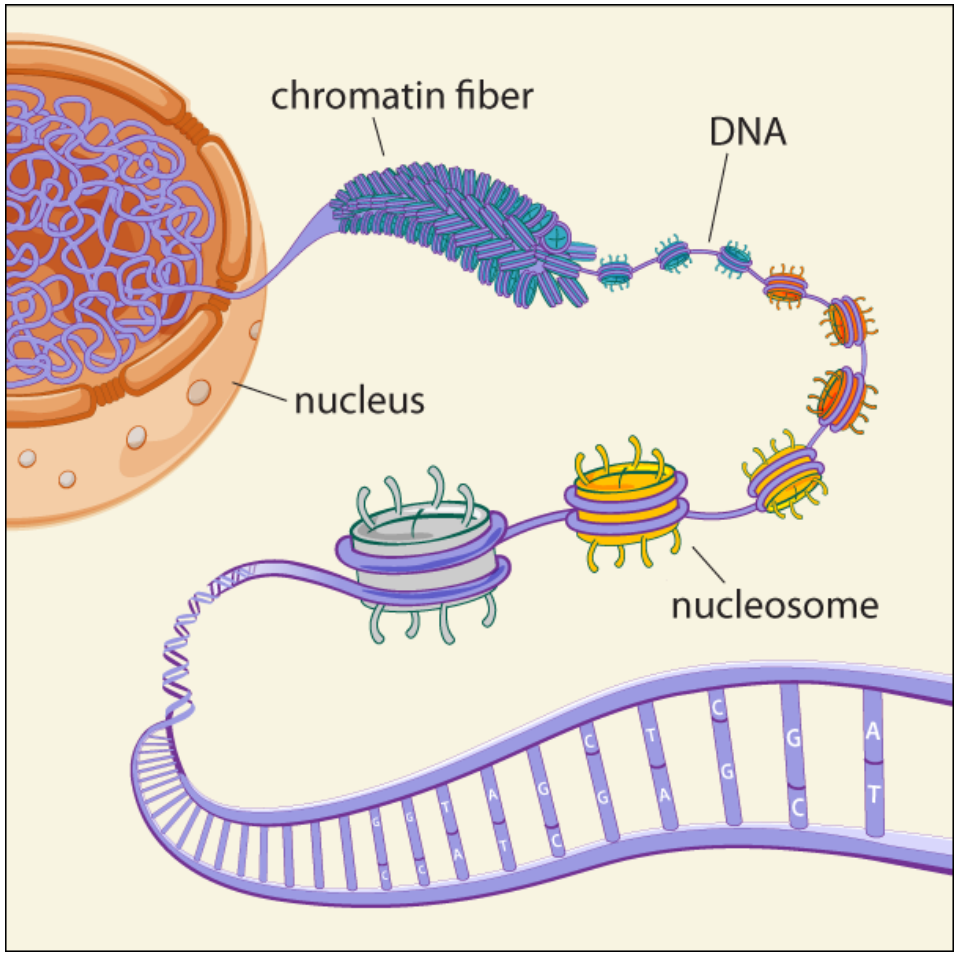
Renata Varnaitė [1], Stuart A MacNeill [2]

BIOTINYLATION BY ANTIBODY RECOGNITION (BAR): MEASURING PPIS FOR ANY PROTEIN IN SITU



Bar, Nat Meth, 2018
Wu, et al. Metab Eng, 2025

DNA-PROTEIN INTERACTIONS



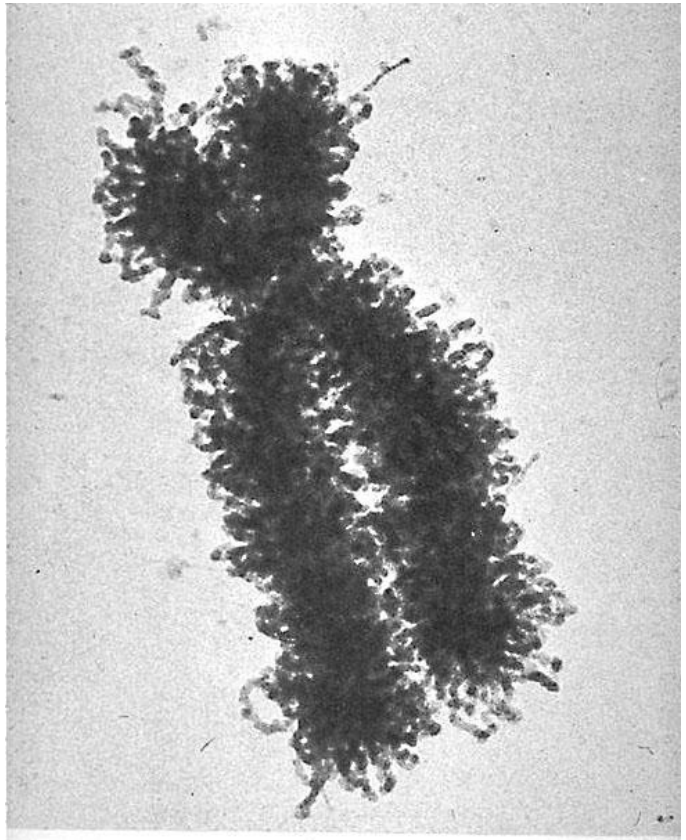
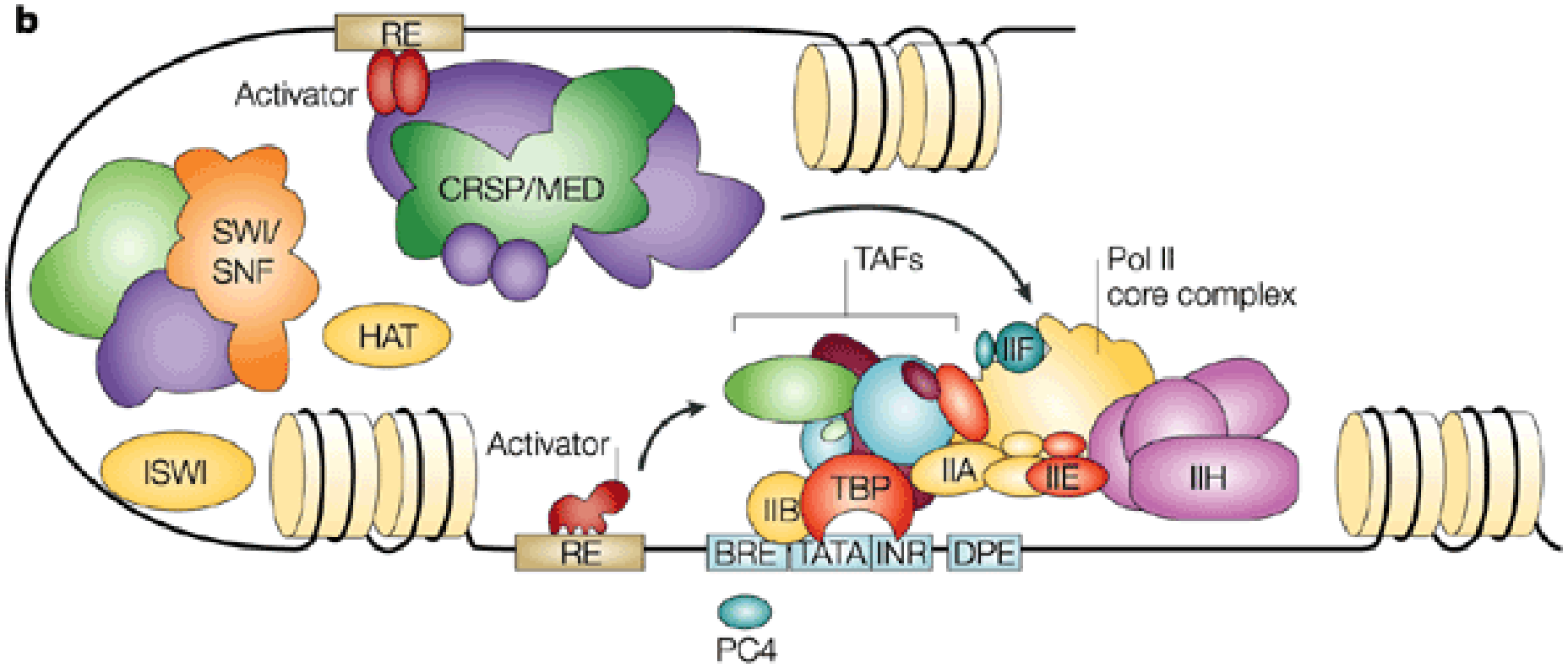


FIGURE 1-14
An electron micrograph of a human chromosome. Chromosome XII from a HeLa cell culture. (Courtesy of Dr. E. Du Praw.)

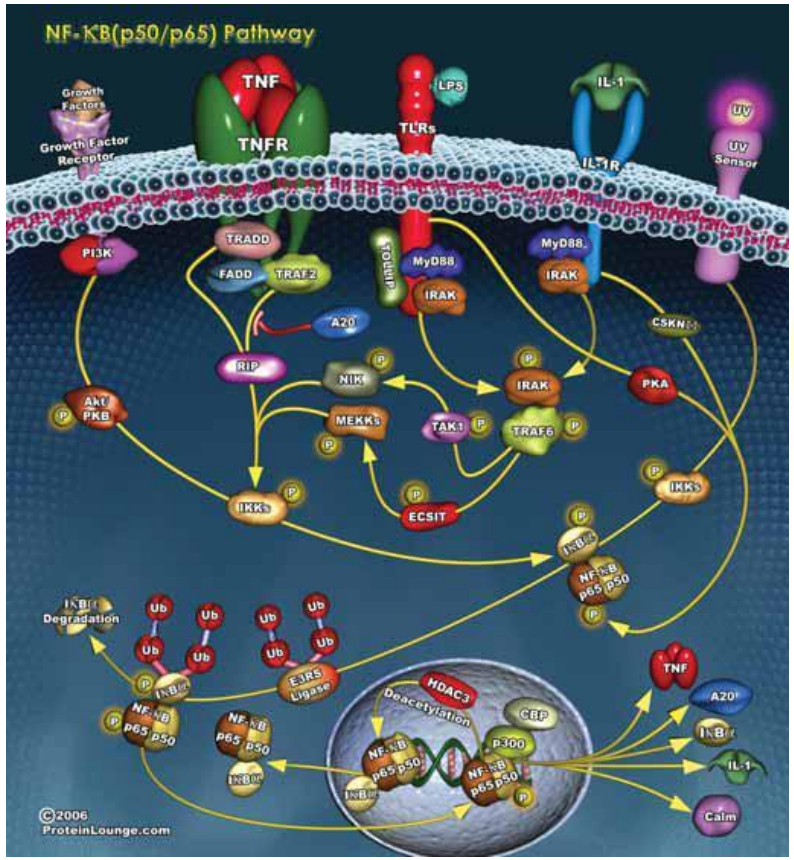
MANY PROTEIN DNA INTERACTIONS AT A EUKARYOTIC PROMOTER

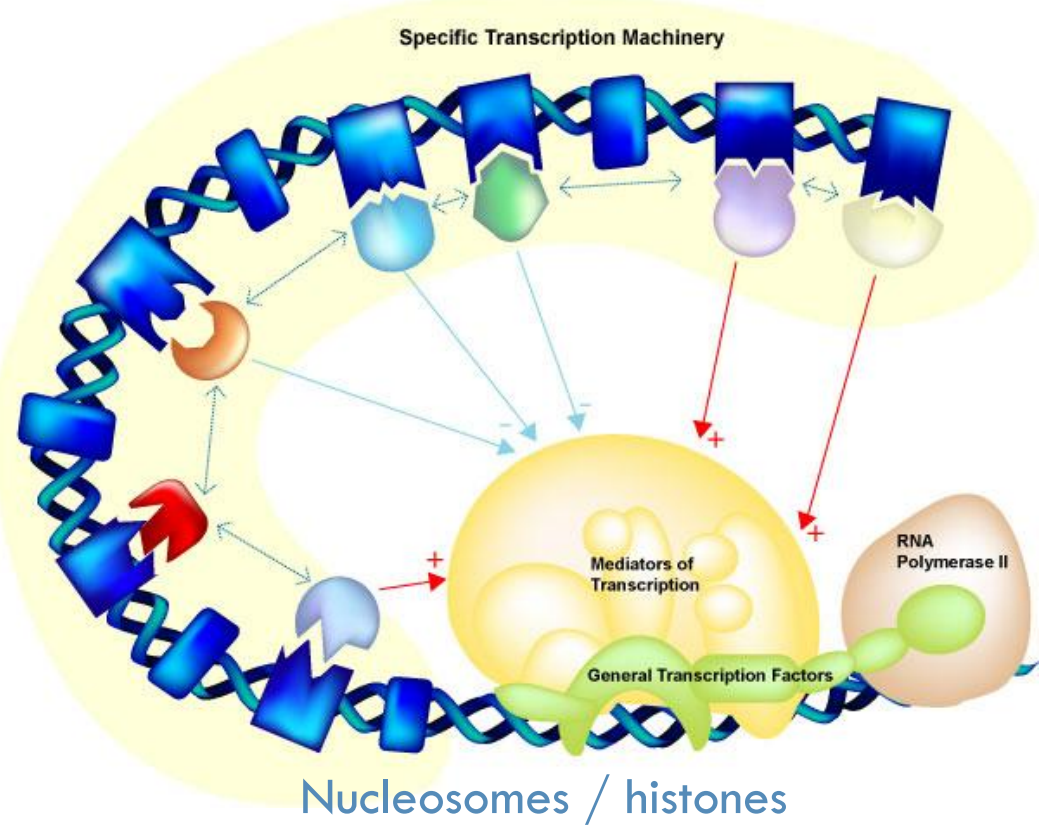


Dylan J. Taatjes, Michael T. Marr & Robert Tjian
Nature Reviews Molecular Cell Biology, 2004

Nature Reviews | Molecular Cell Biology

DNA-PROTEIN INTERACTIONS OF INTEREST





Nucleosomes / histones

Transcription factors

- Cis and trans regulators

RNA polymerase

DNA modifying enzymes

SOME EXAMPLE TECHNOLOGIES

Study the interaction of specific DNA sequences and specific proteins

- Gel shift assay
- DNase footprint
- Microfluidic approaches
- Protein-binding microarrays
- SELEX
- Yeast 1-hybrid
- Chip-Chip/Chip-seq
- ChIP-exo – base pair resolution of protein binding

Study whole-genome-scale DNA-protein interactions or transcription sites

- DNase-seq
- FAIRE
- Pol2-seq

Chromatin structure and high-level interactions

- 3C / 5C techniques

Enhancers, etc.

- STARR-seq - high throughput determination of enhancer strengths

EARLY TECHNIQUES

Gel shift assay

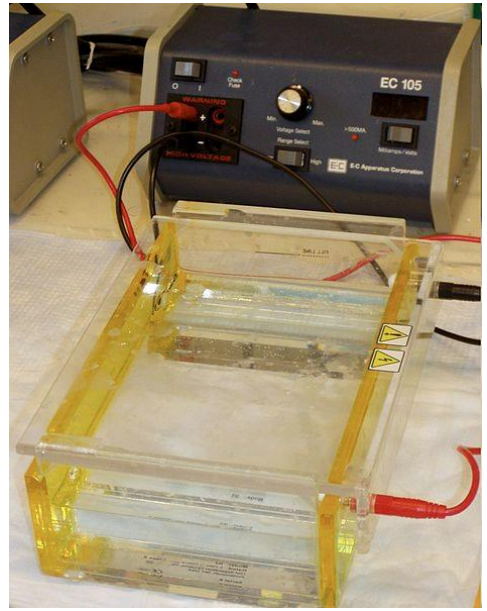
Detecting DNA-protein interaction by electrophoresis

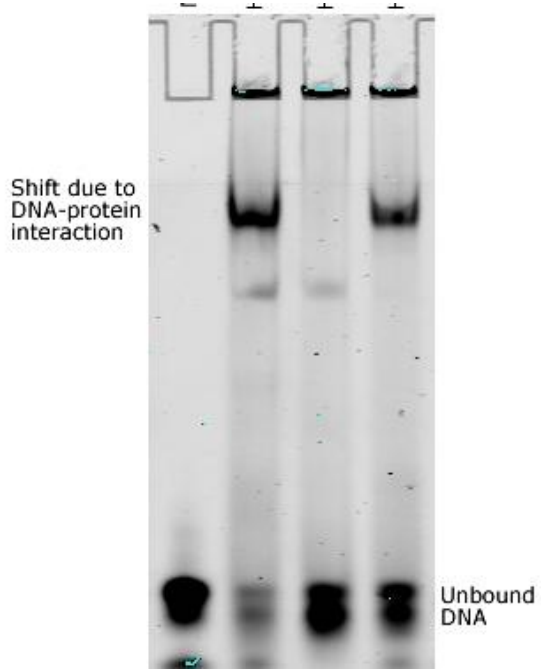
Incubate DNA and proteins of interest together

- Require that you know DNA sequence of interest a priori

Load mixture onto electrophoresis gel

Mobility of DNA bound by proteins migrate slower in a gels





DNASE FOOTPRINTING

Label DNA sequences of interest

Incubate with protein of interest with DNA

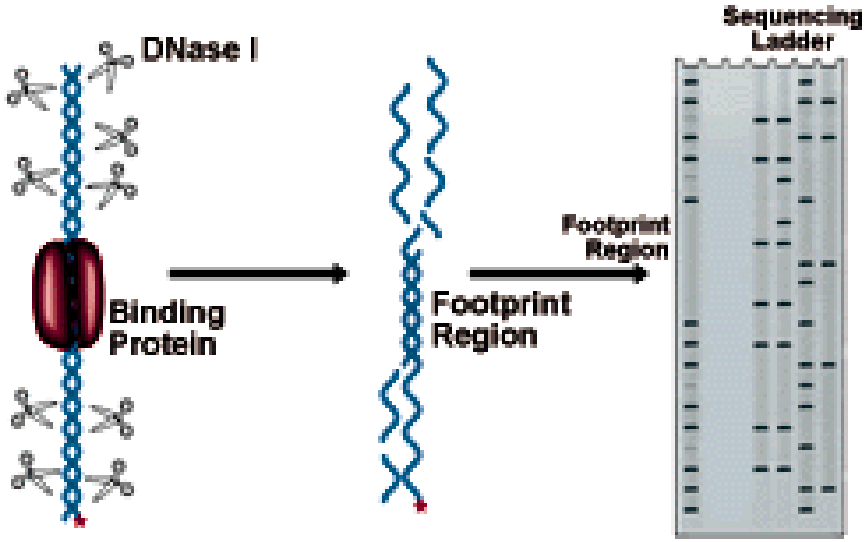
Subject samples to DNaseI

Expect random cuts

Protein binding will protect DNA from DNaseI

Run on gel to detect

Test different concentration of the protein

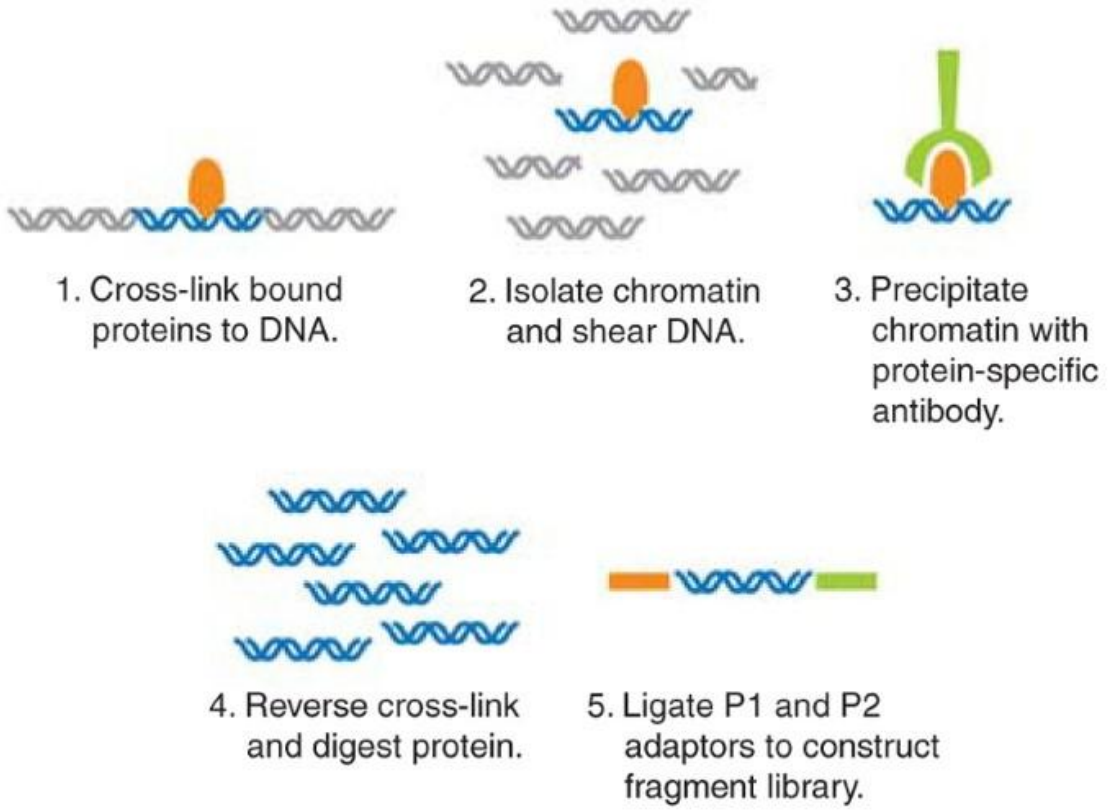


HIGH-THROUGHPUT *IN VIVO*: CHIP-SEQ

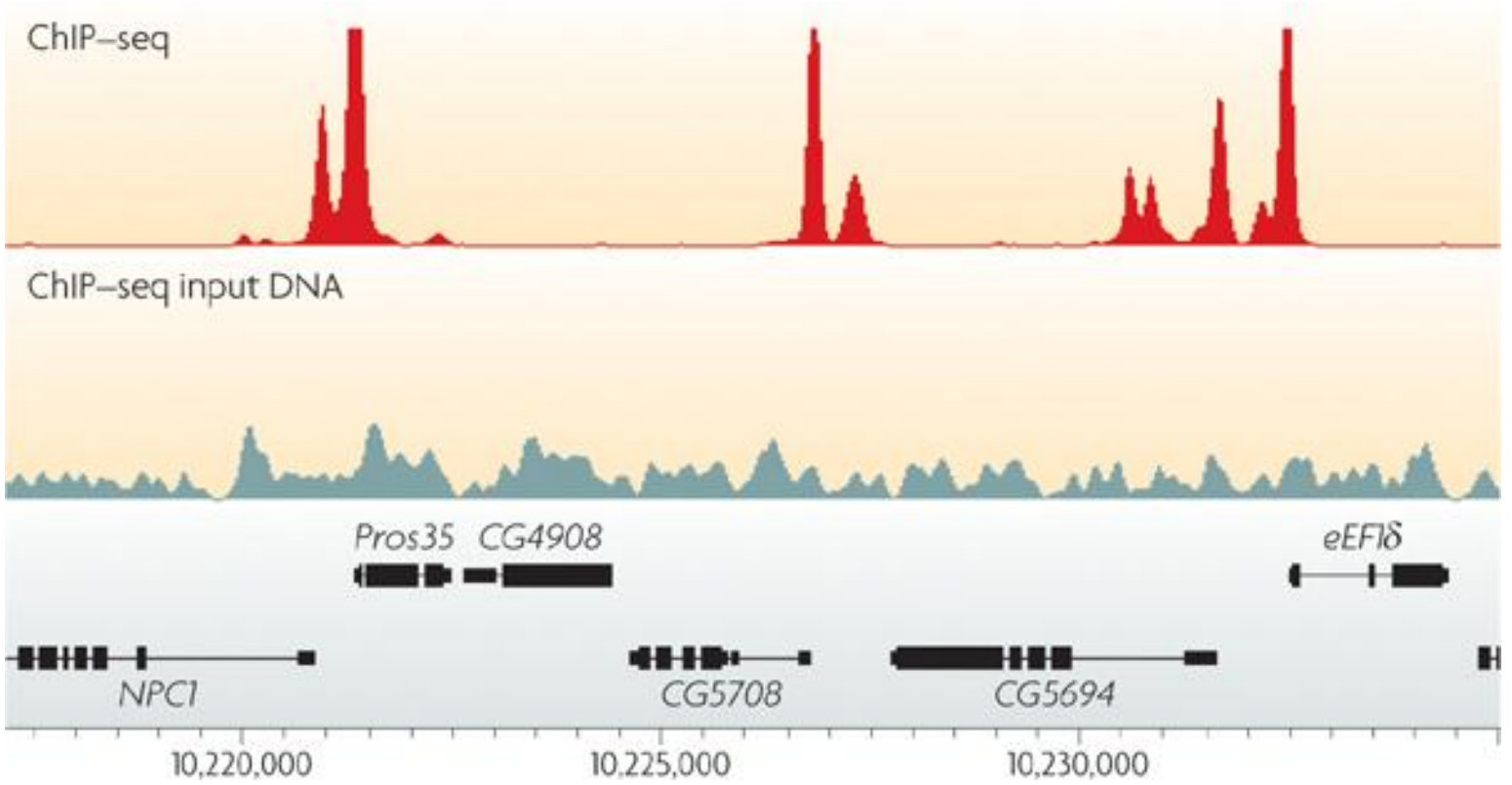
Combining **Ch**romatin-**i**mmuno**p**recipitation and DNA microarray (**chip**)

Enabling detection the genome-wide binding events of a protein of interest

Require antibody against the protein of interest



CHIP-SEQ… HIGHER RESOLUTION

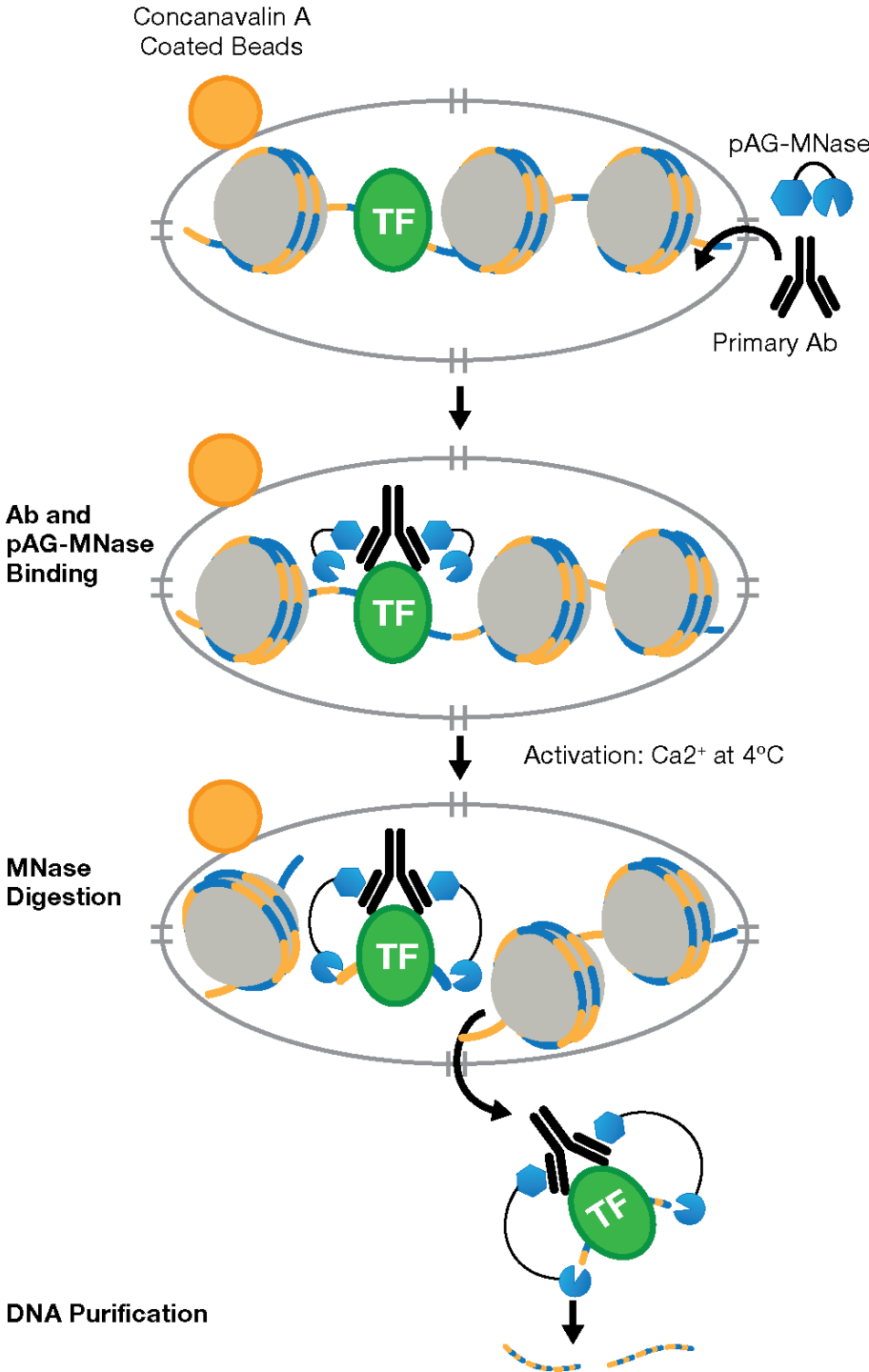


CUT&RUN

Cleavage Under Targets & Release Using Nuclease

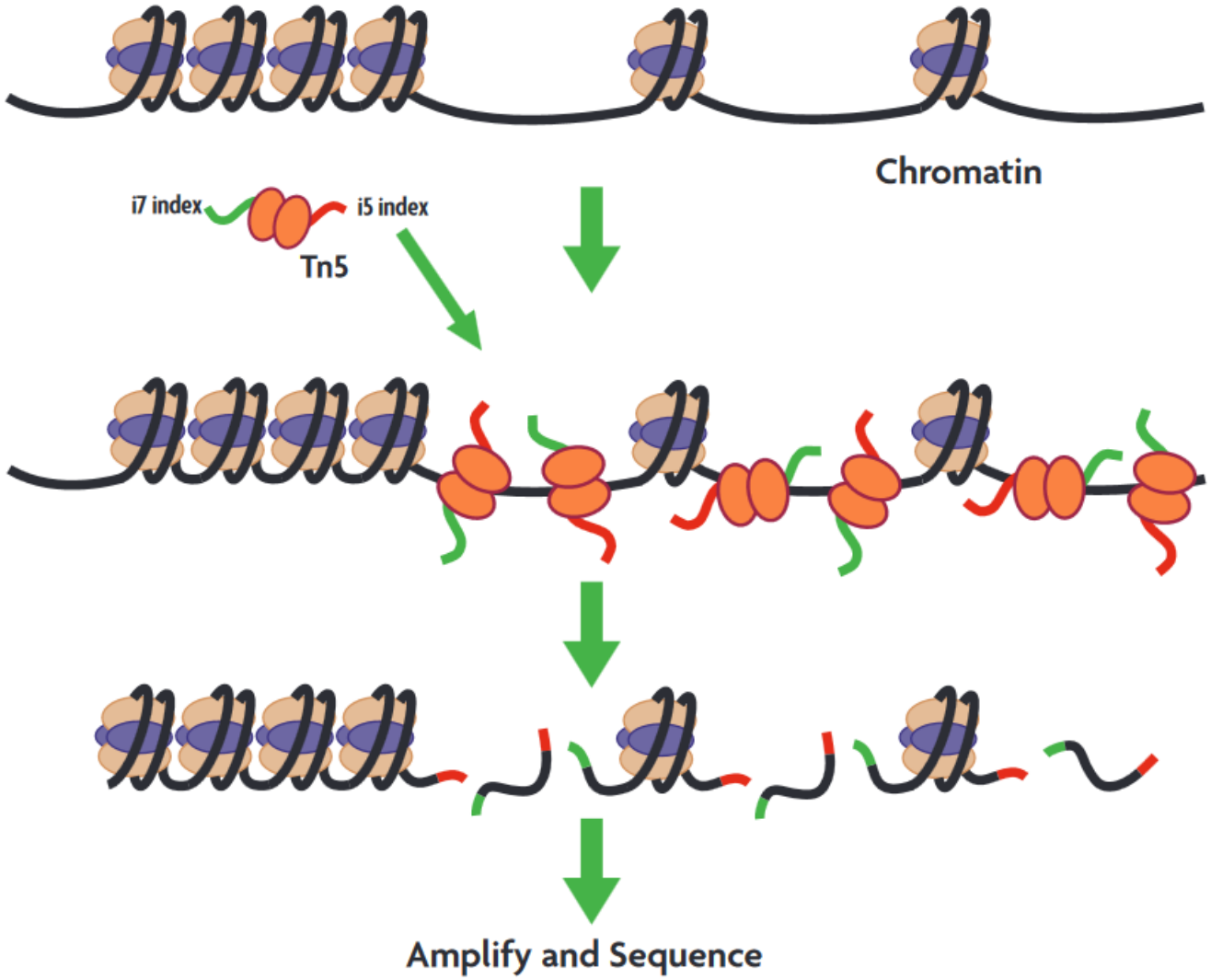
Less sample material and less sequencing depth req.

No cross-linking



Nice video/primer: https://www.activemotif.com/applications-cut-and-run

Finding open chromatin: ATAC-seq

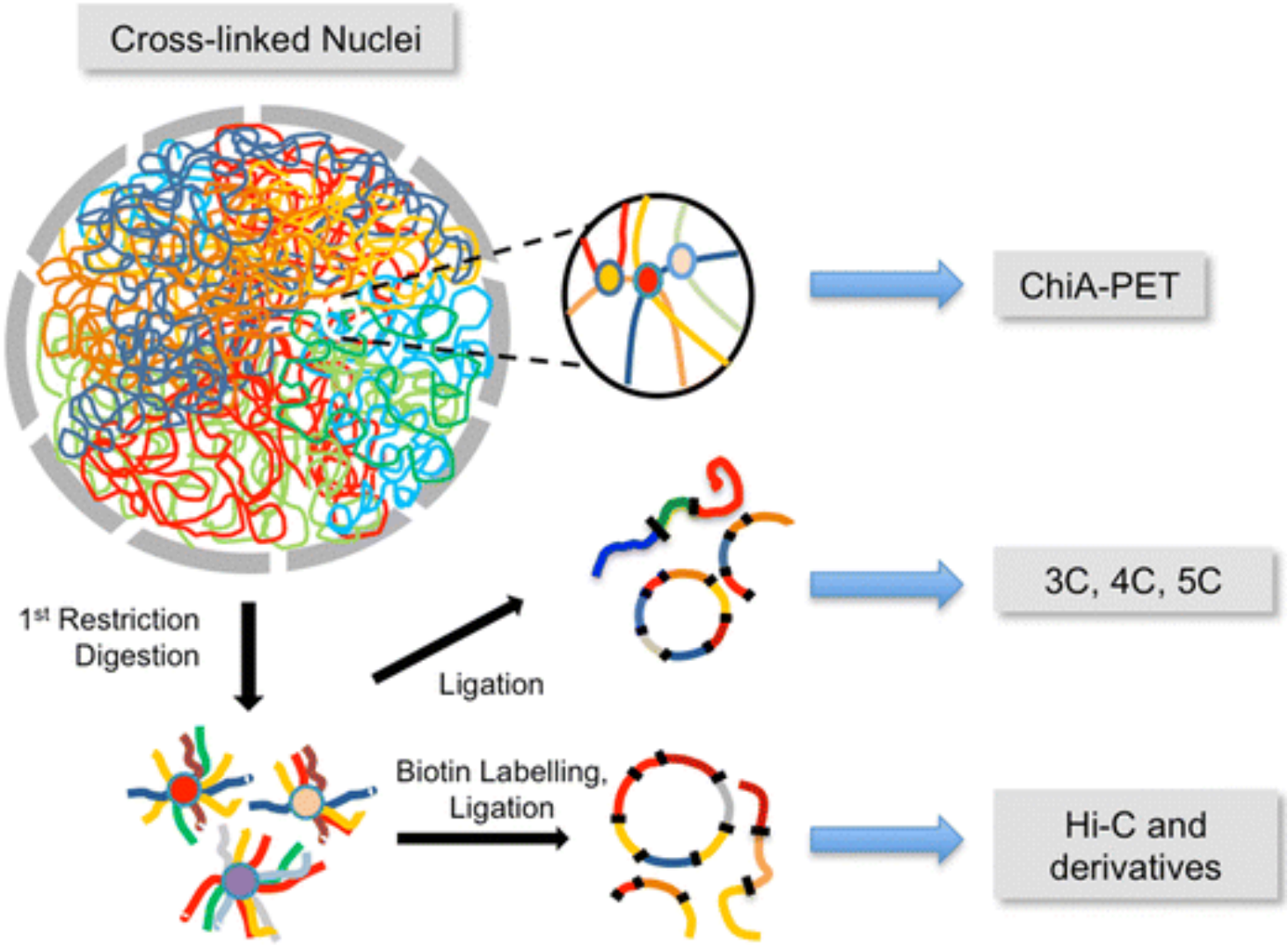


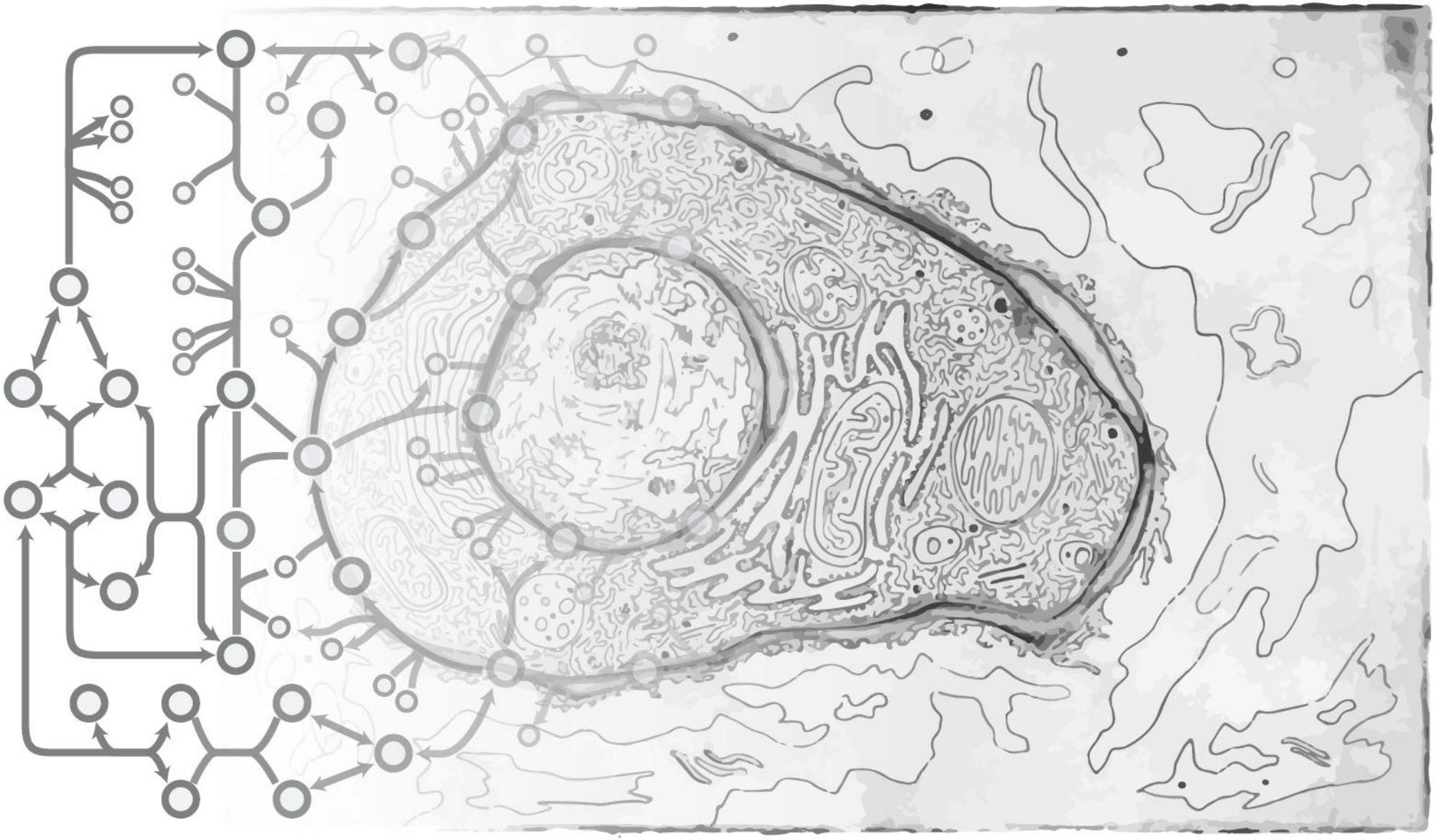
CHROMATIN HIGH-LEVEL STRUCTURES

Chromosome Configuration Capture (3C); Circularized Chromosome Conformation Capture (4C); Chromosome Configuration Capture Carbon Copy (5C)

Detecting long distance interactions between fragments of chromosome, e.g., interaction of enhancers and transcription machinery

Using formaldehyde to cross-link interacting proteins and DNA fragments, followed by sequencing and mapping to genome



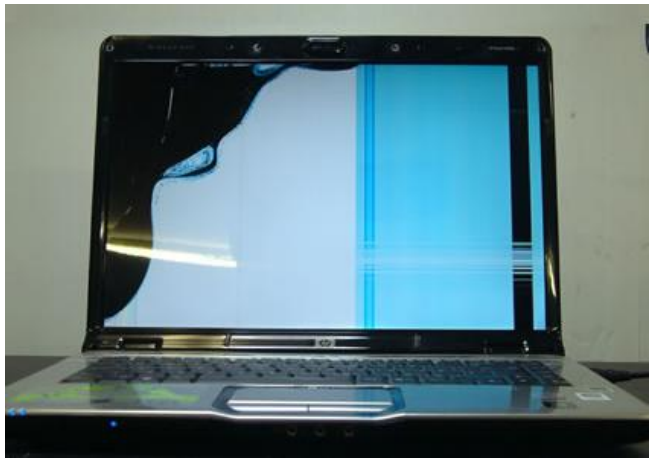


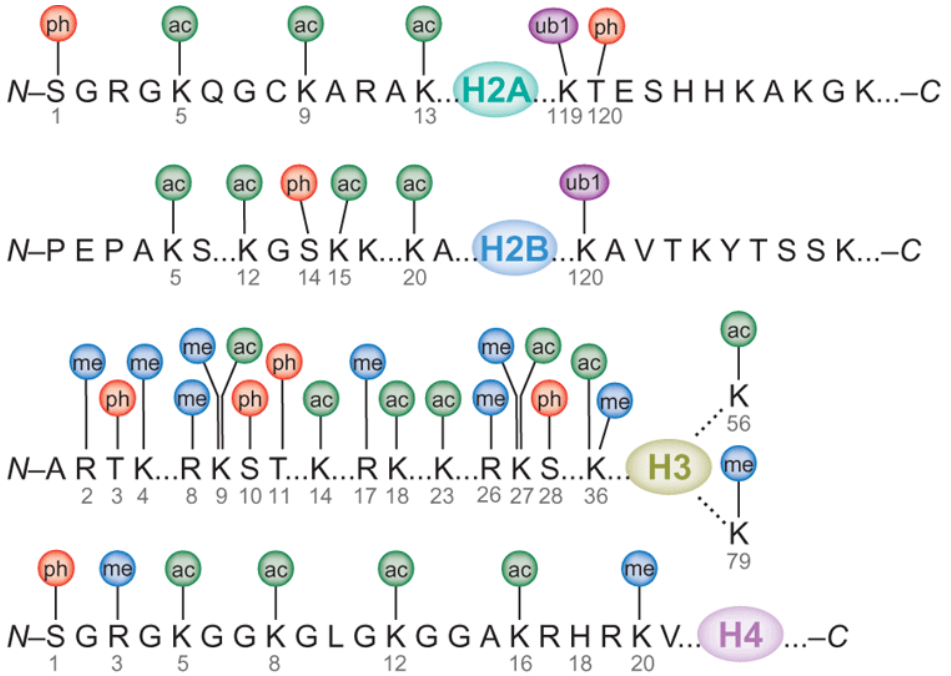
CHEMICAL MODIFICATION OF THE PARTS LIST

COVALENT MODIFICATIONS

Changes in chemical bonding

Changes function





PROTEOMIC ANALYSIS OF POST-TRANSLATIONAL MODIFICATIONS

Post-translational modifications (PTMs)

- Covalent processing events that change the properties of a protein
  - proteolytic cleavage
  - addition of a modifying group to one or more amino acids
- Determine its activity state, localization, turnover, interactions with other proteins
- Mass spectrometry and other biophysical methods can be used to determine and localize potential PTMs
  - However, PTMs are still challenging aspects of proteomics with current methodologies

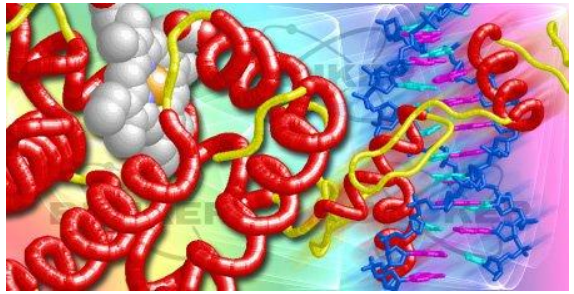
COMPLEXITY OF THE PROTEOME

Protein processing and modification comprise an important third dimension of information, beyond those of DNA sequence and protein sequence.

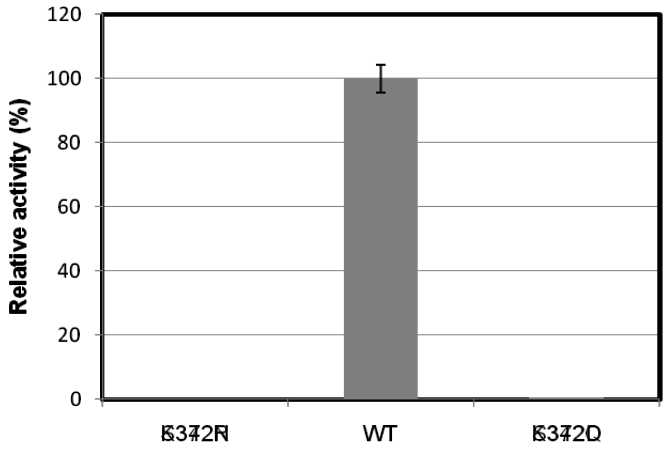
Complexity of the human proteome is far beyond the more than 30,000 human genes.

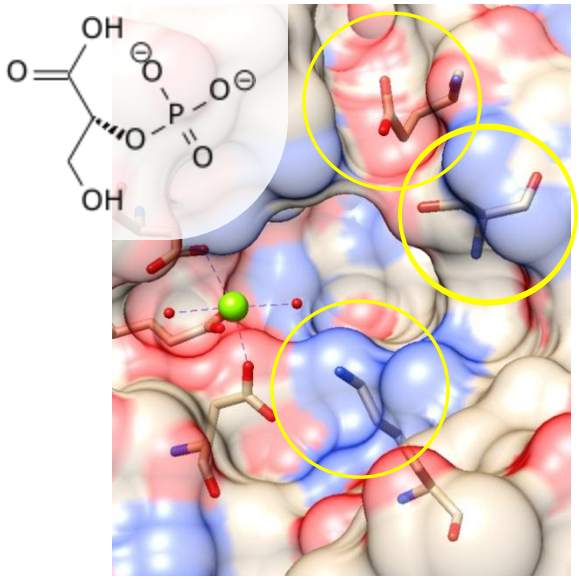
The thousands of component proteins of a cell and their post-translational modifications may change with the cell cycle, environmental conditions, developmental stage, and metabolic state.

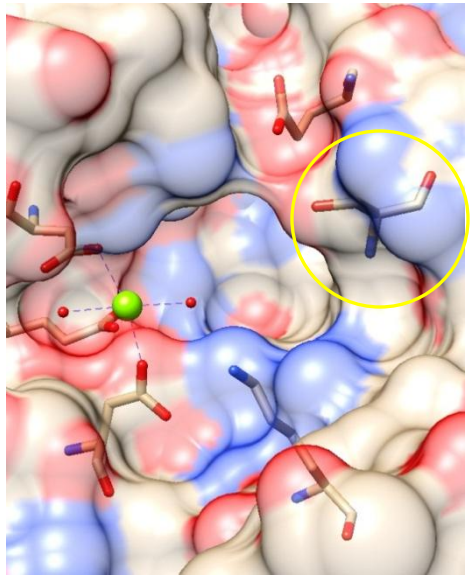
Proteomic approaches that advance beyond identifying proteins to elucidating their post-translational modifications are needed.



PTMS REGULATE ACTIVE SITES: ENOLASE

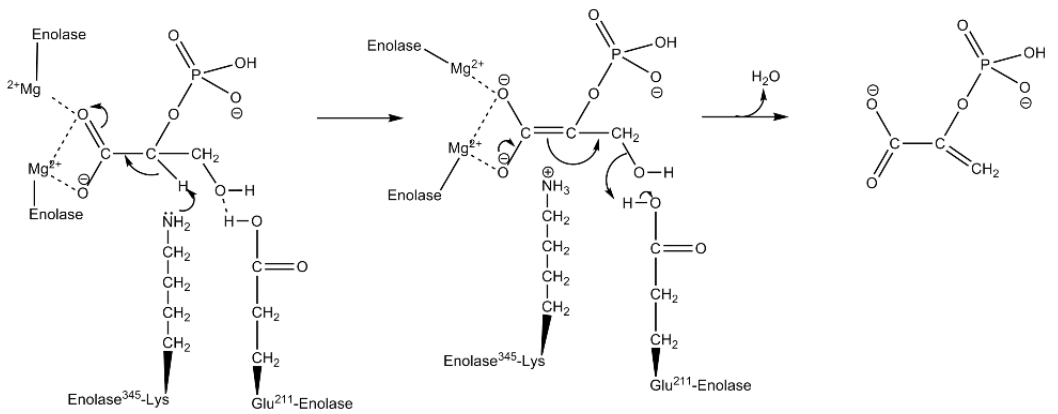






Acetylation occurs on active site lysine

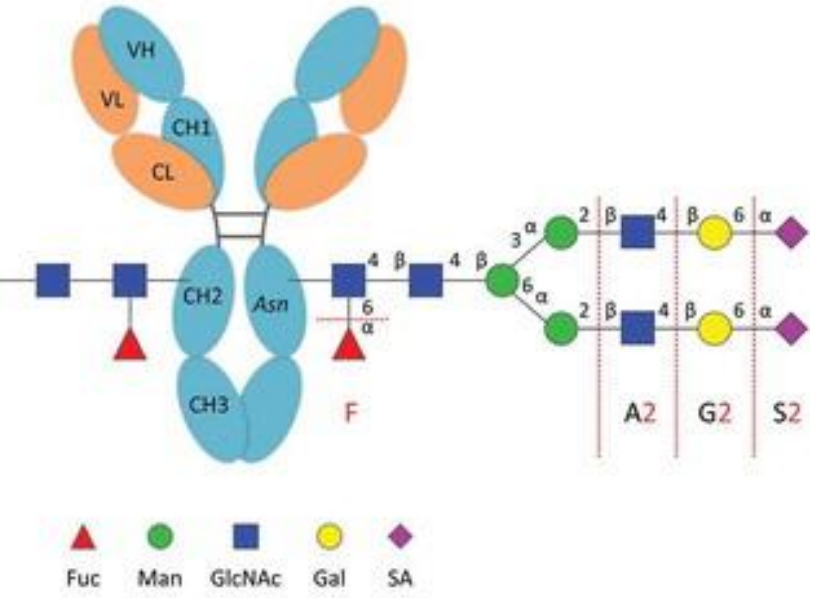
Phosphorylation charge blocks active site

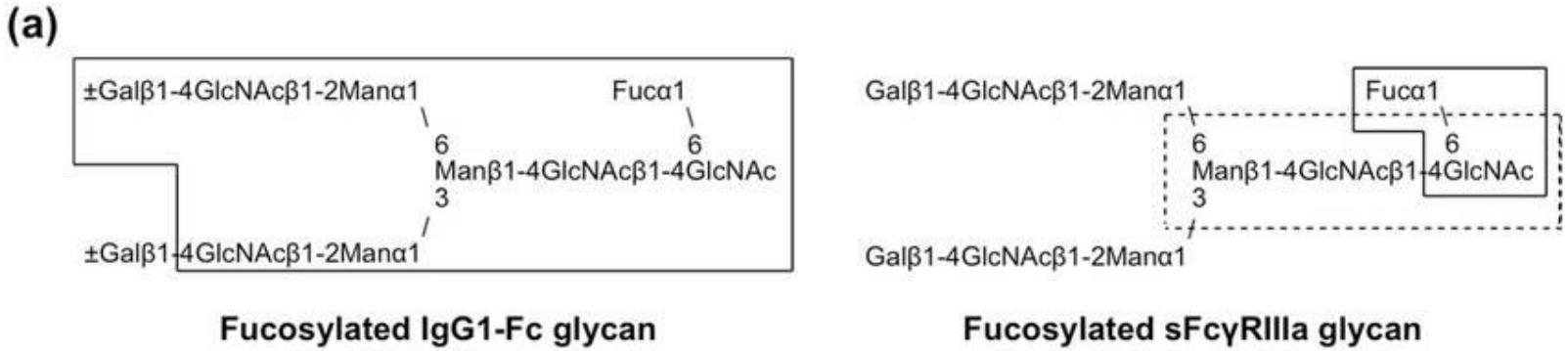


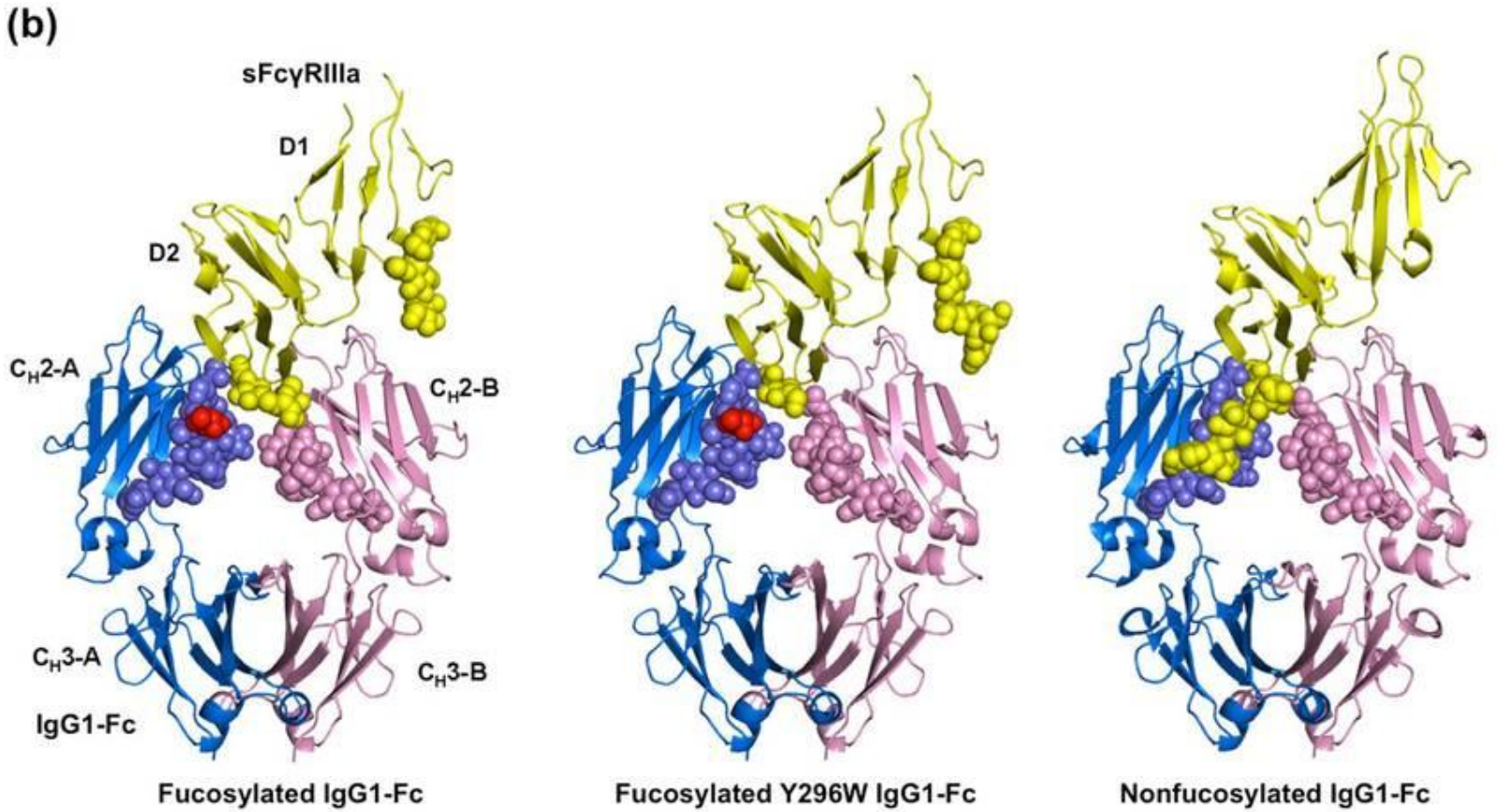
2-phosphoglycerate phosphoenolpyruvate

Brunk, et al PNAS 2018

PTMS REGULATE PROTEIN INTERACTIONS







EXAMPLE: PHOSPHORYLATION

Analysis of the entire complement of phosphorylated proteins in cells: "phosphoproteome"

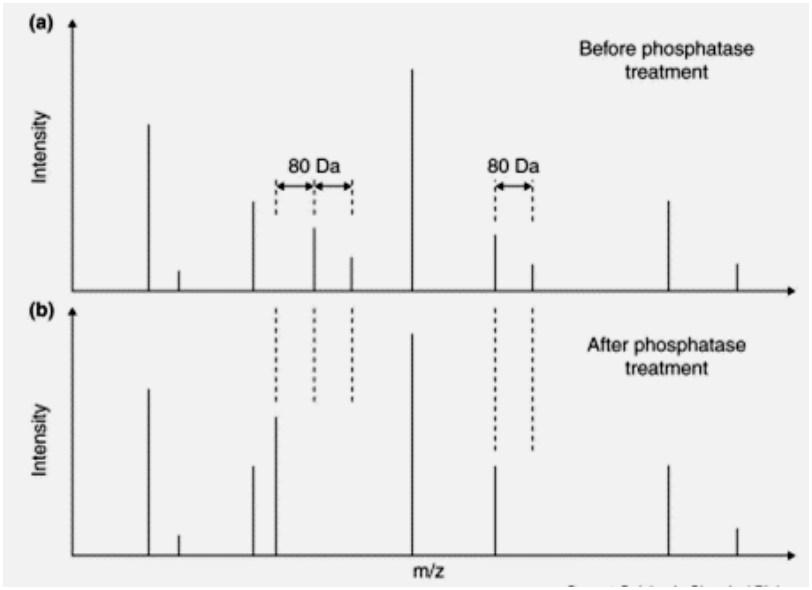
Qualitative and quantitative information regarding protein phosphorylation important

Important in many cellular processes

- signal transduction, gene regulation, cell cycle, apoptosis

Most common sites of phosphorylation: Ser, Thr, Tyr

- MS can be used to detect and map locations for phosphorylation
  - MW increase from addition of phosphate group
  - treatment with phosphatase allows determination of number of phosphate groups
  - digestion and tandem MS allows for determination of phosphorylation sites



OTHER COVALENT MODIFICATIONS

a. Phosphorylation - involved in regulation

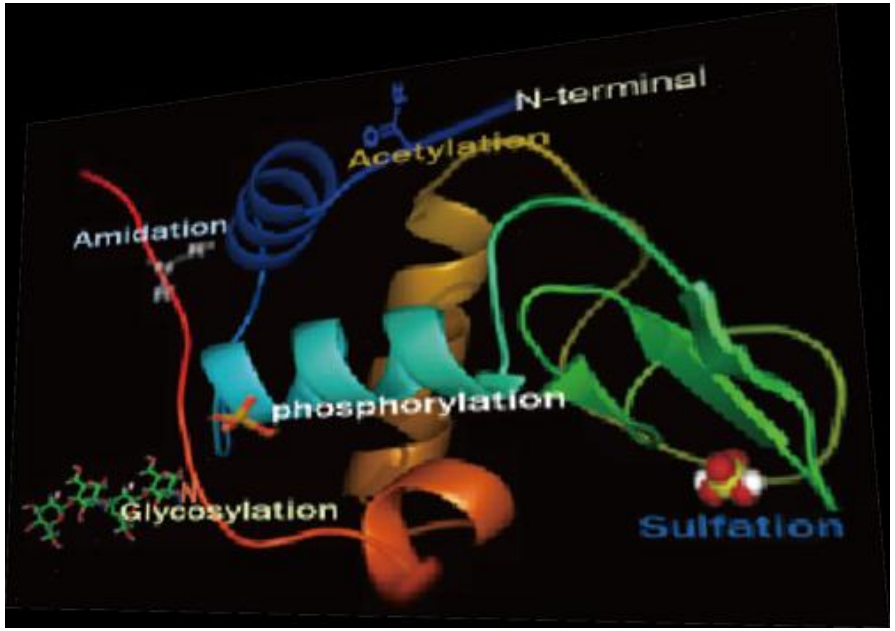
b. Glycosylation - common outside cells (especially biopharma)

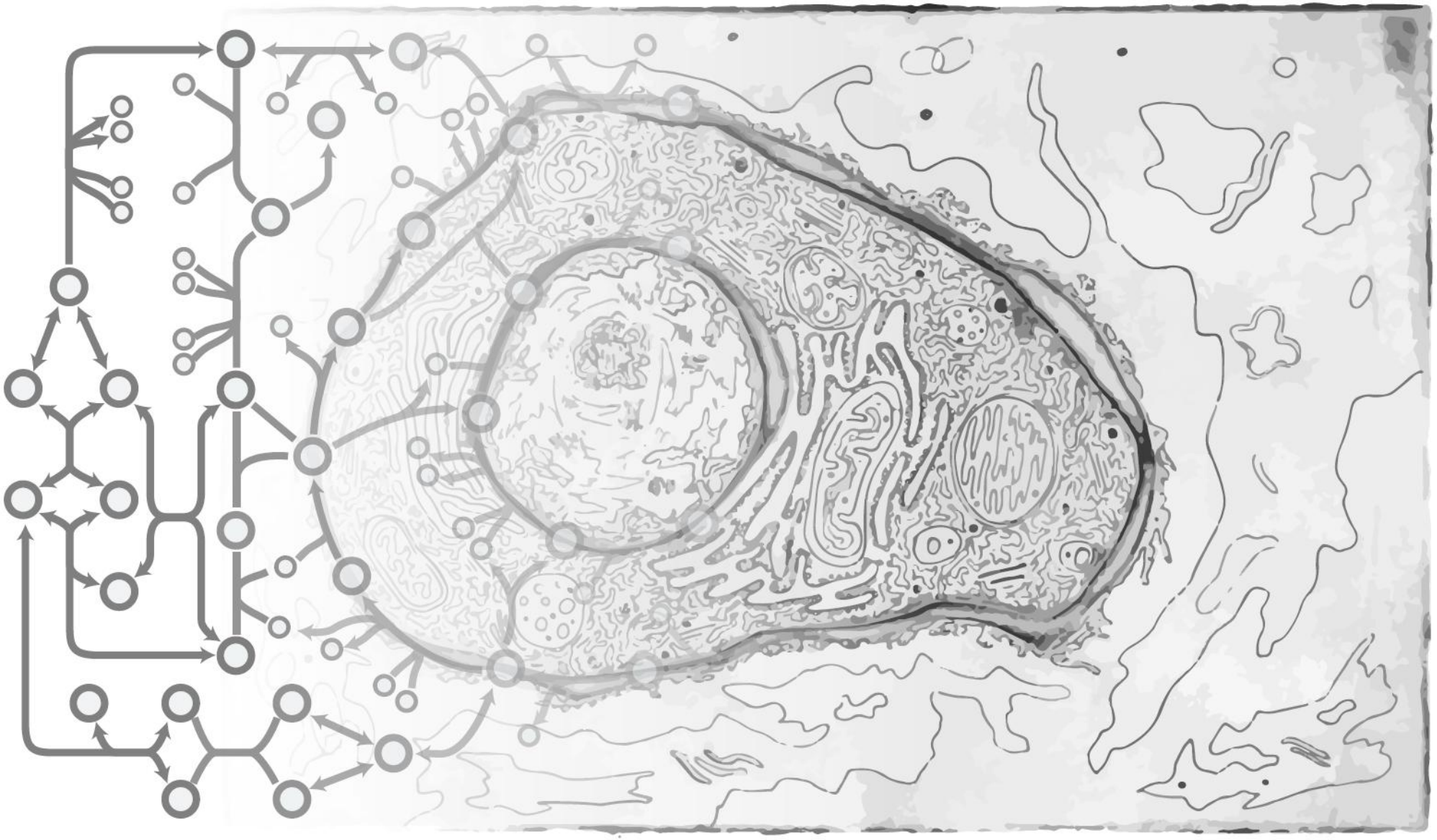
c. Lipidation - attach to membrane

d. Nitrosylation - involved in regulation

e. Acetylation - common on first aa.

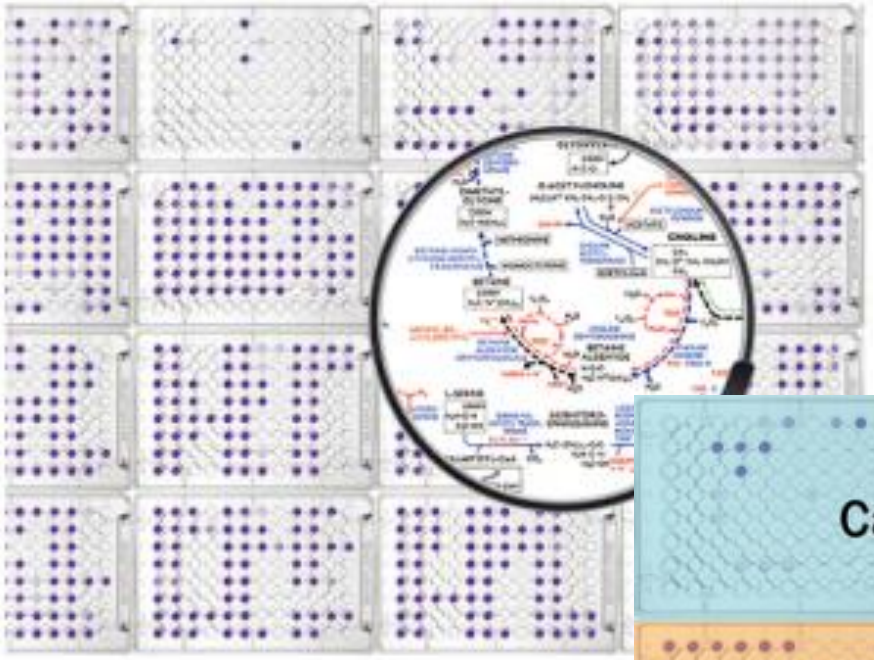
f. Ubiquitination - for degradation





INTERVENTIONAL OMICS: POKING AT CELLS AND SEEING WHAT THEY DO

GROWTH SCREENS ON DIFFERENT MEDIA COMPONENTS: BIOLOG



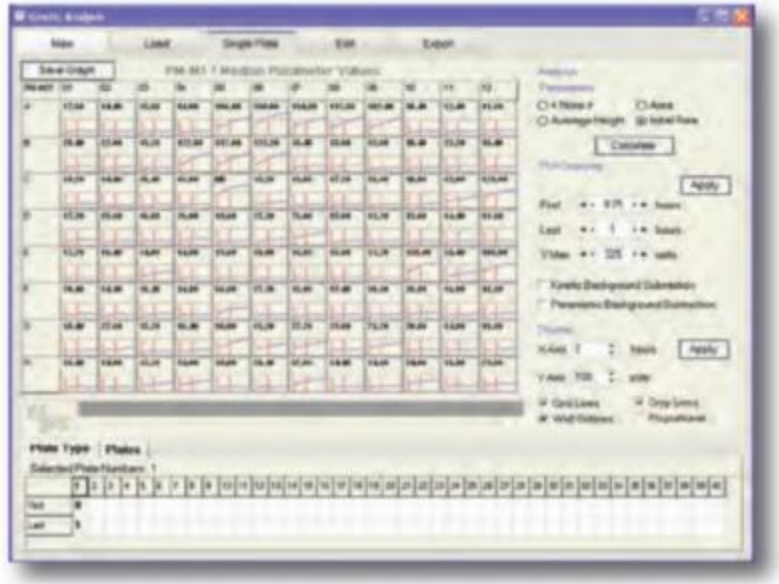
367
Carbon-Energy and Nitrogen Substrates

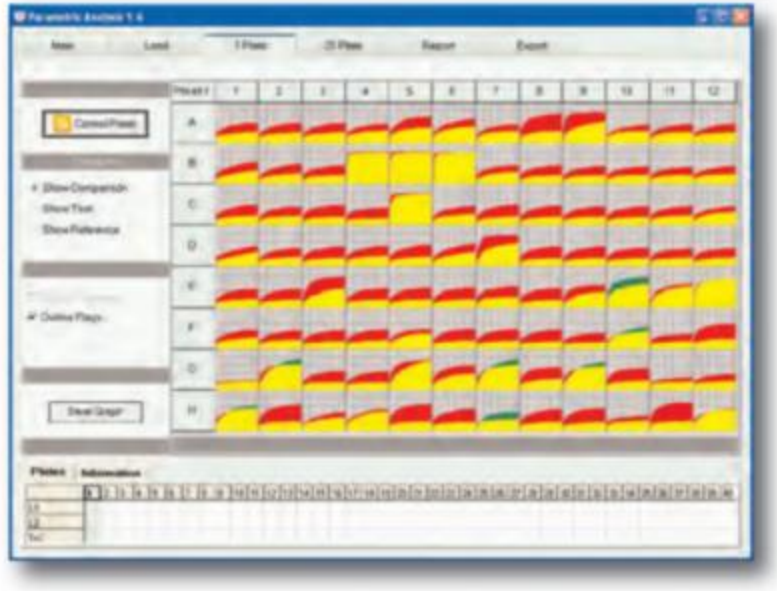
22
Titrated Ions

45
Titrated Hormones and Cytol

92
Titrated Anti-Cancer Drugs

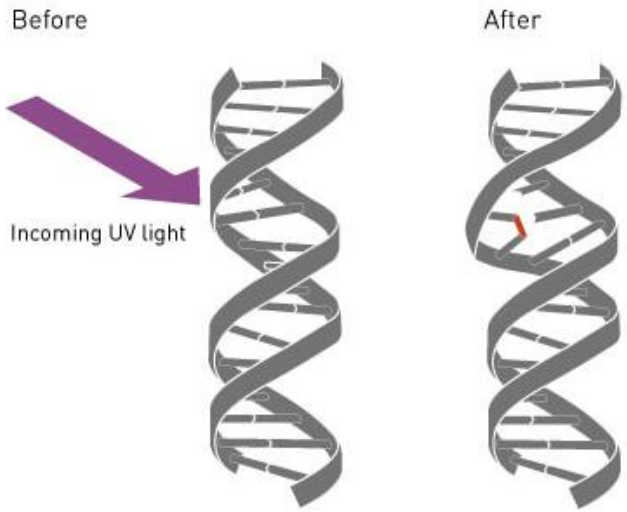
8
Carbon-Energy Substrates





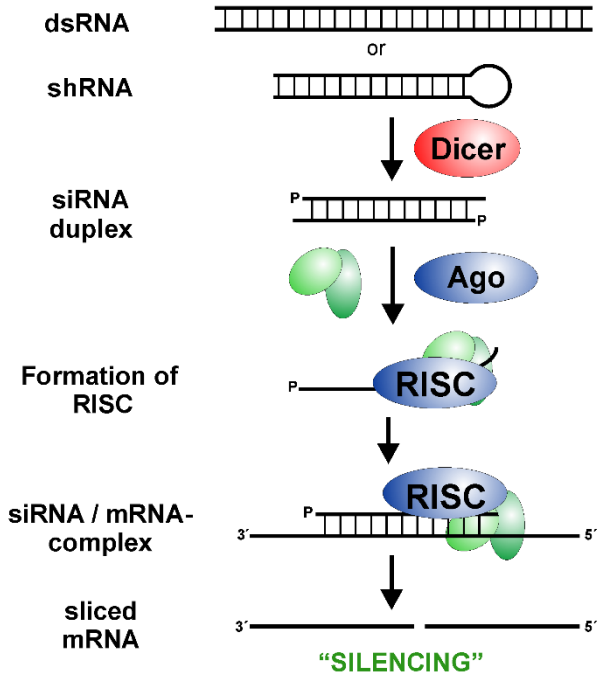
FORWARD SCREENING IN MAMMALIAN CELLS

RANDOM MUTAGENESIS



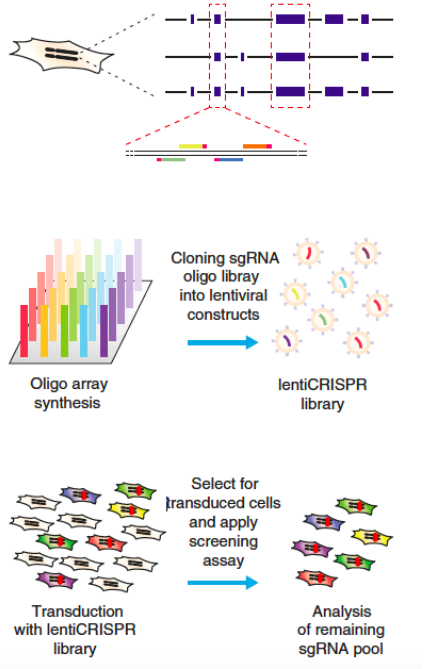
https://www.singerinstruments.com/

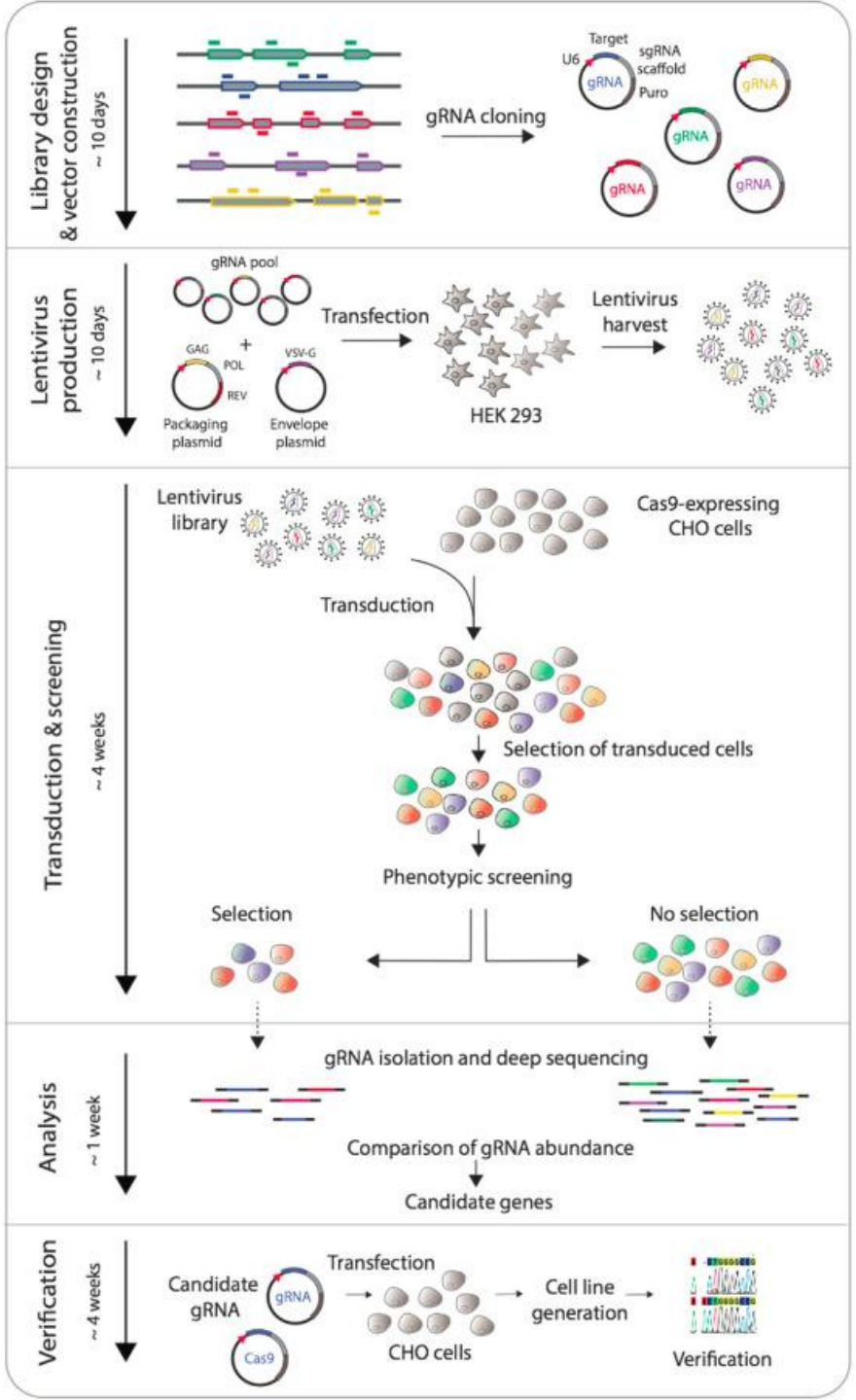
SiRNA SCREENS



http://www.gene-quantification.de/si-rna.html

CRISPR SCREENS





CRISPR SCREENS

Allow the identification of genes associated with a trait of interest

Goal: Make pool of cells, each with 1 trackable edit

Logic: If you can isolate/enrich/select "good" cells you can ID what edit is responsible

La Cour Karottki, Met Eng, 2021

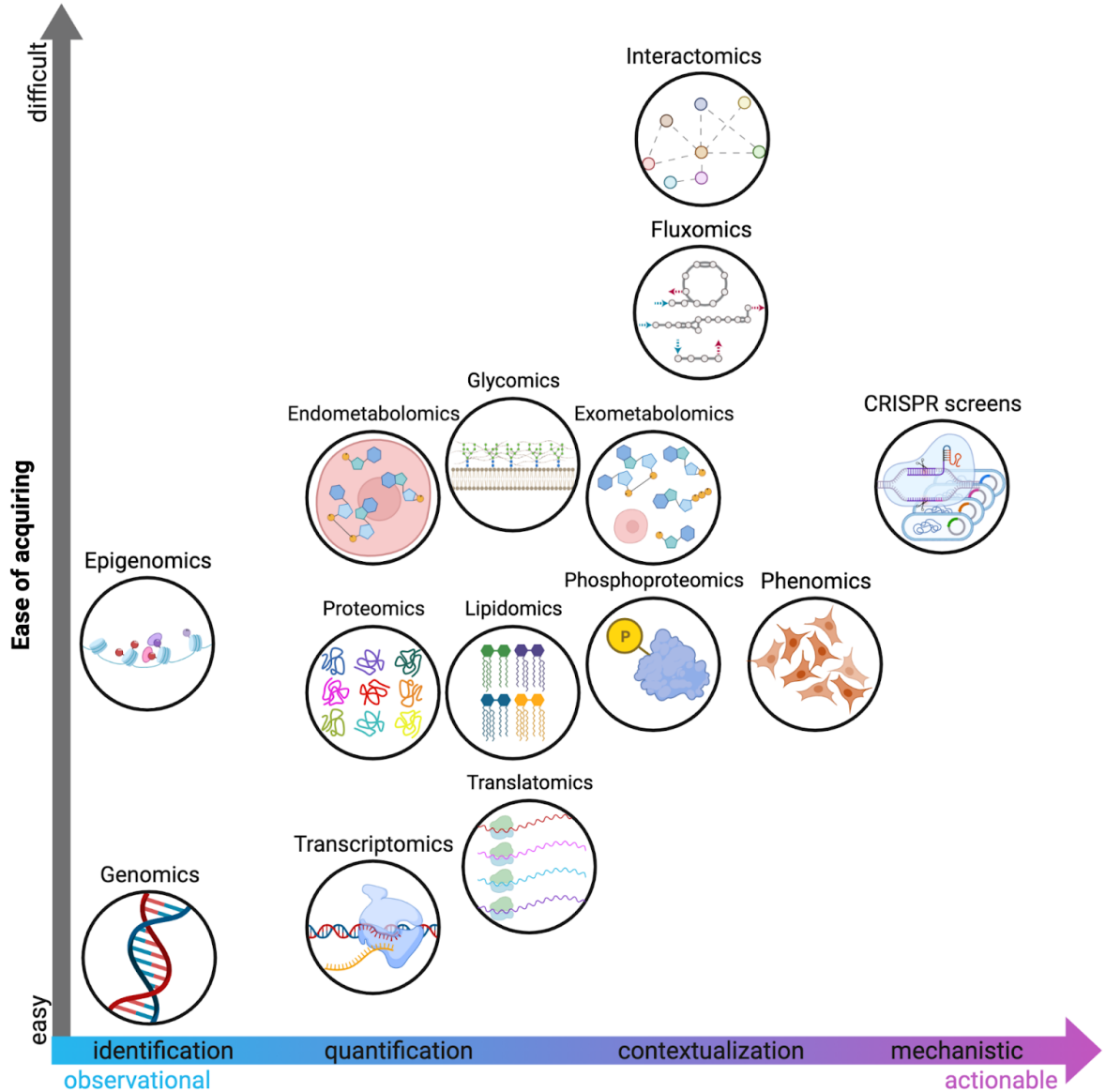
ACTIVITY: "ACTIONABLE" OMICS

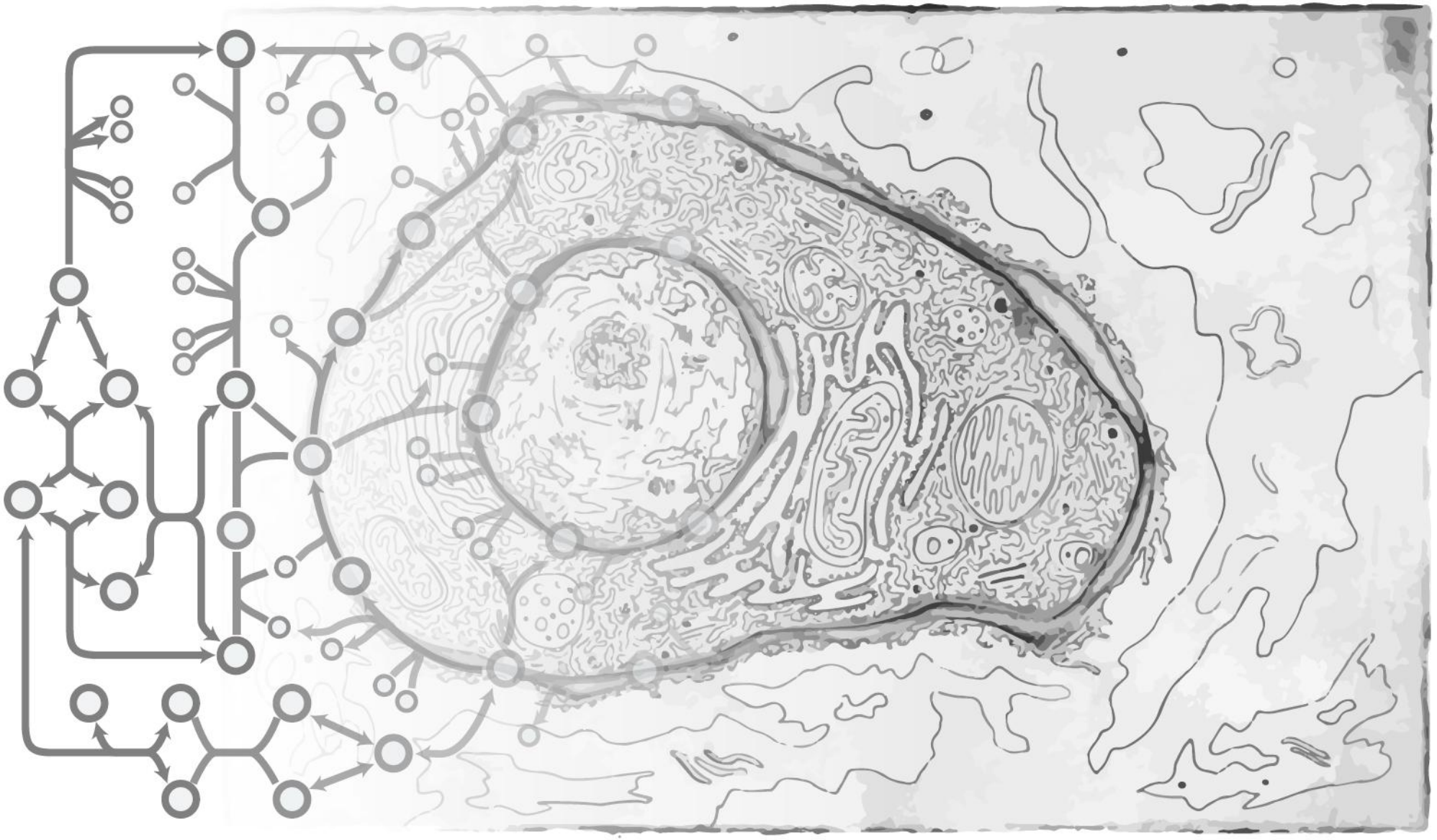
In this lecture we've discussed a number of different omics:

- Genomics
- Transcriptomics
- Proteomics
- Metabolomics
- Interactomics: protein-protein interactions
- Interactomics: protein-DNA interactions
- Epigeomics
- Biolog media screens
- CRISPR screens

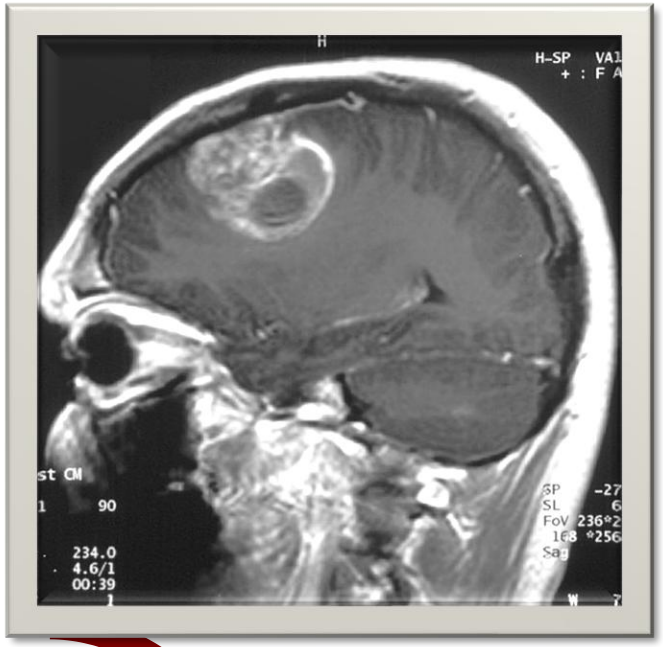
Form groups of 2-3 students

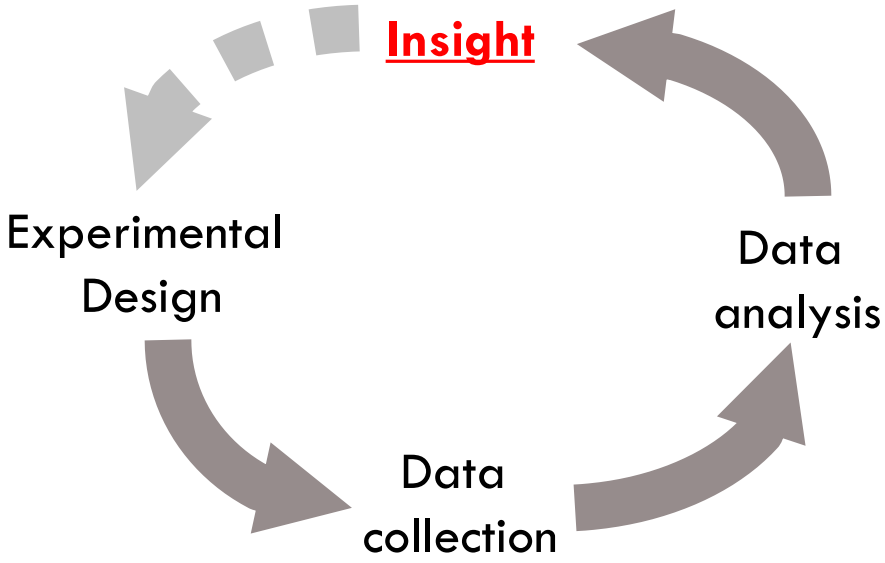
Take each of these and order them with regard to how "actionable" they could be for cell line development and engineering. Describe what the data could tell, and how you may act on the results.

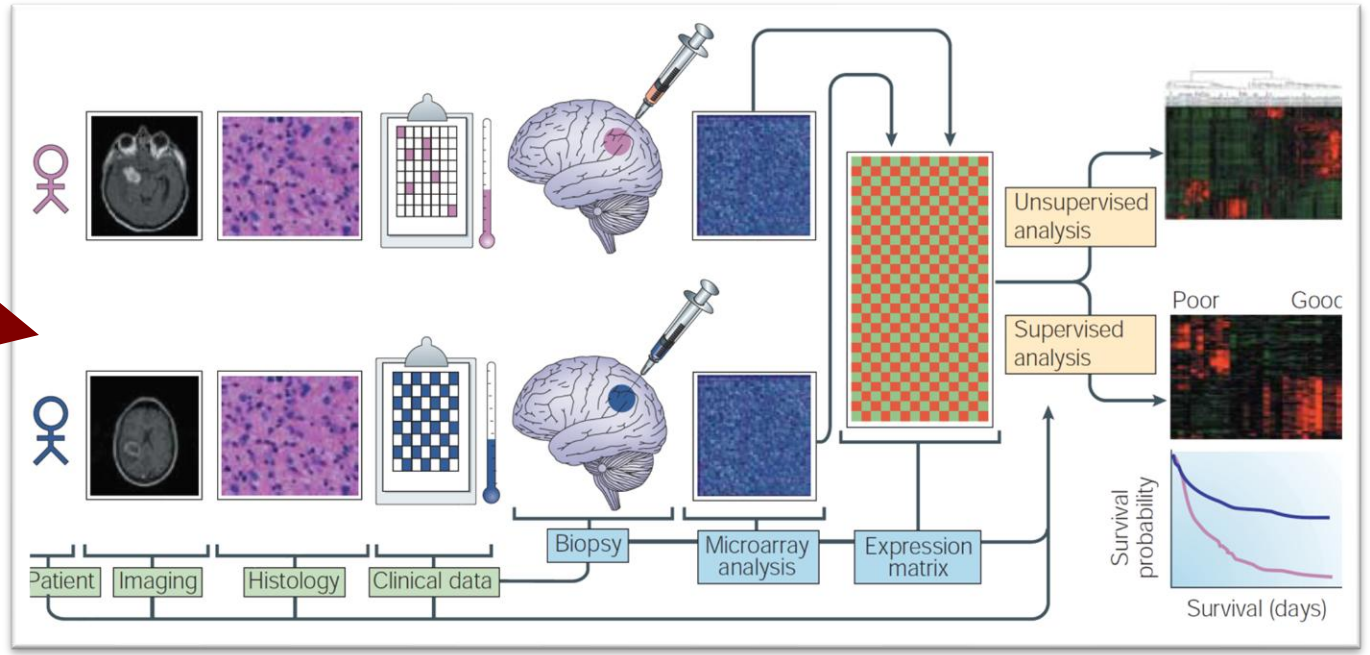

Ease of acquiring
difficult
easy
Interactomics
Fluxomics
Glycomics
Endometabolomics
Exometabolomics
CRISPR screens
Epigenomics
Proteomics
Lipidomics
Phosphoproteomics
Phenomics
P
Translatomics
Transcriptomics
Genomics
identification
quantification
contextualization
mechanistic
observational
actionable



MAKING SENSE OF OMICS DATA







Mischel et al, 2004

```
[54271] "37226_at"      "37232_at"
[54273] "37254_at"      "37278_at"
[54275] "37384_at"      "37408_at"
[54277] "37424_at"      "37425_g_at"
[54279] "37433_at"      "37462_i_at"
[54281] "37512_at"      "37547_at"
[54283] "37549_g_at"    "37566_at"
[54285] "37577_at"      "37586_at"
[54287] "37590_g_at"    "37652_at"
[54289] "37793_r_at"    "37796_at"
[54291] "37802_r_at"    "37831_at"
[54293] "37860_at"      "37872_at"
[54295] "37892_at"      "37943_at"
[54297] "37950_at"      "37953_s_at"
[54299] "37965_at"      "37966_at"
[54301] "37986_at"      "37996_s_at"
[54303] "38037_at"      "38043_at"
[54305] "38069_at"      "38149_at"
[54307] "38157_at"      "38158_at"
[54309] "38241_at"      "38269_at"
[54311] "38290_at"      "38340_at"
[54313] "38398_at"      "38447_at"
[54315] "38487_at"      "38521_at"
[54317] "38671_at"      "38691_s_at"
[54319] "38703_at"      "38707_r_at"
[54321] "38710_at"      "38766_at"
[54323] "38892_at"      "38918_at"
[54325] "38964_r_at"    "39248_at"
[54327] "39249_at"      "39313_at"
[54329] "39318_at"      "39402_at"
[54331] "39548_at"      "39549_at"
[54333] "39582_at"      "39650_s_at"
[54335] "396_f_at"      "39705_at"
[54337] "39729_at"      "39763_at"
[54339] "39817_s_at"    "39835_at"
[54341] "39854_r_at"    "39891_at"
[54343] "39966_at"      "40016_g_at"
[54345] "40020_at"      "40093_at"
[54347] "40148_at"      "40149_at"
[54349] "40189_at"      "40225_at"
[54351] "40255_at"      "40273_at"
[54353] "40284_at"      "40359_at"
[54355] "40420_at"      "40446_at"
[54357] "40465_at"      "40472_at"
[54359] "40489_at"      "40524_at"
[54361] "40560_at"      "40562_at"
[54363] "40569_at"      "40612_at"
[54365] "40640_at"      "40665_at"
[54367] "40687_at"      "40829_at"
[54369] "40837_at"      "40850_at"
[54371] "41037_at"      "41047_at"
[54373] "41113_at"      "41160_at"
[54375] "41220_at"      "41329_at"
[54377] "41386_i_at"    "41387_r_at"
[54379] "41397_at"      "41469_at"
[54381] "41512_at"      "41553_at"
[54383] "41577_at"      "41644_at"
[54385] "41657_at"      "41660_at"
[54387] "41856_at"      "41858_at"
[54389] "42361_g_at"    "43427_at"
[54391] "43511_s_at"    "43544_at"
```

WHAT DO YOU DO WITH ALL THESE DATA?

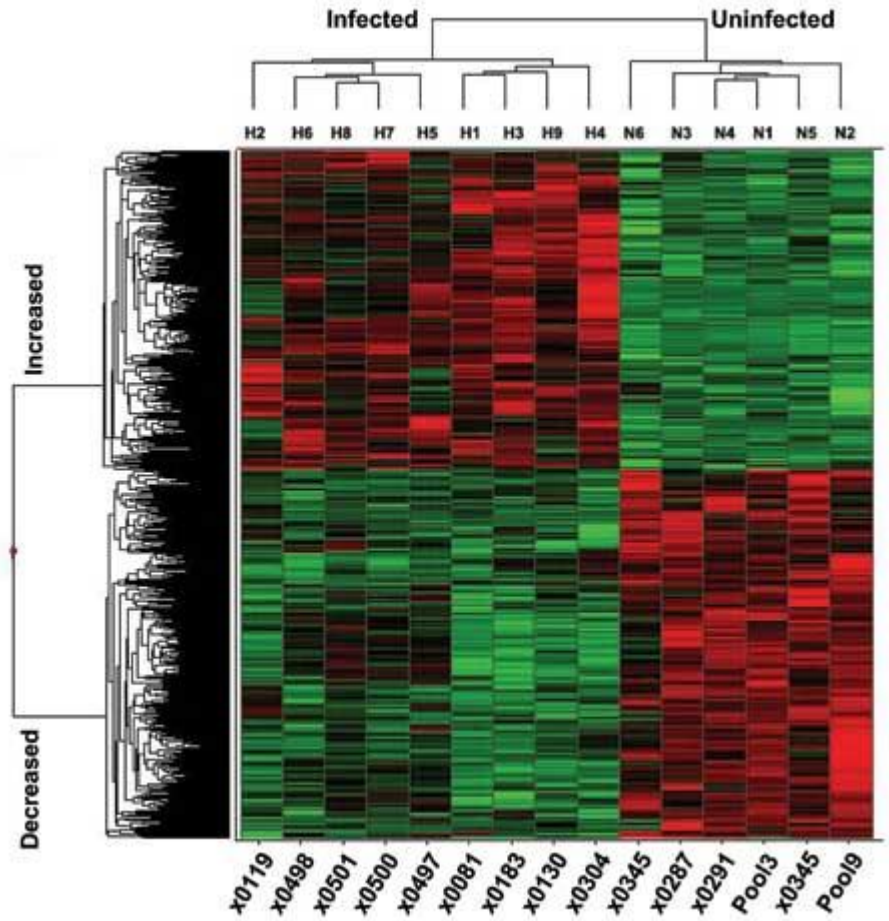
Exploratory analysis

- Enrichment
- Principal component analysis
- Clustering

Each analysis can return a set of “important” genes

Making sense of it requires various levels of gene annotation

- Individual genes
- Groups
- Pathways



WHAT DO ALL OF THESE GENES DO?!

Fortunately, lots of scientists have an idea…

Even better, others have collected all of this information

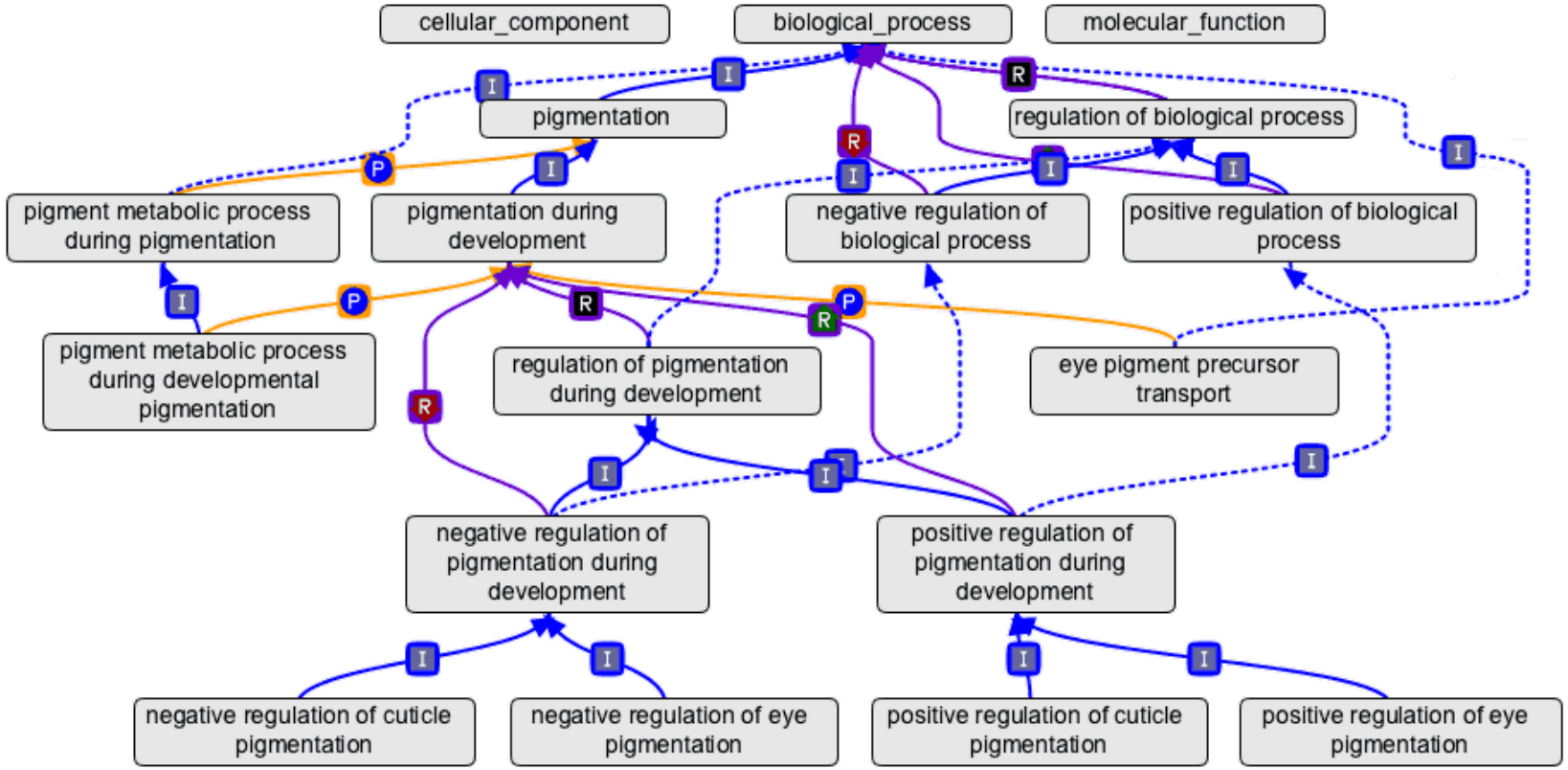
- Find commonality among the gene
  - Common molecular functions (GO)
  - Common biological processes (GO)
  - Common cellular components (GO)
  - Common pathways
  - Interact with common genes
  - Common sequences / molecular structures
  - Regulated by common Transcription Factors
  - Targeted by common microRNAs
  - Involved in the same disease
- Generate new hypothesis based on the commonality of genes in your set

Structured controlled vocabularies (ontologies) that describe relationships between gene products and their associated biological roles

- *cellular components* : the parts of a cell or its extracellular environment
- *molecular functions*: activities, such as catalytic or binding activities, that occur at the molecular level (e.g. catalytic activity, Toll receptor binding)
- *biological processes:* series of events accomplished by one or more ordered assemblies of molecular functions (e.g. signal transduction, pyrimidine metabolic process)

GO structure

- Directed Acyclic Graph(DAG)
- Child terms are more specialized
- Child can have more than one parent



Not secure | geneontology.org

GENEONTOLOGY Unifying Biology

About Ontology Annotations Downloads Help

ALLIANCE of GENOME RESOURCES FOUNDING MEMBER

Current release 2022-09-19: 43,335 GO terms | 7,493,159 annotations
1,483,687 gene products | 5,257 species (see statistics)

THE GENE ONTOLOGY RESOURCE

The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.

The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

Search GO term or Gene Product in AmiGO ...

Any Ontology Gene Product

GO Enrichment Analysis

Powered by PANTHER

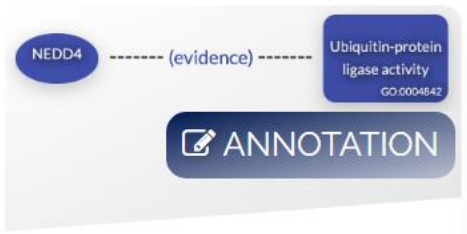
Your gene IDs here...

biological process

Homo sapiens Examples Launch

Hint: can use UniProt ID/AC, Gene Name, Gene Symbols, MOD IDs

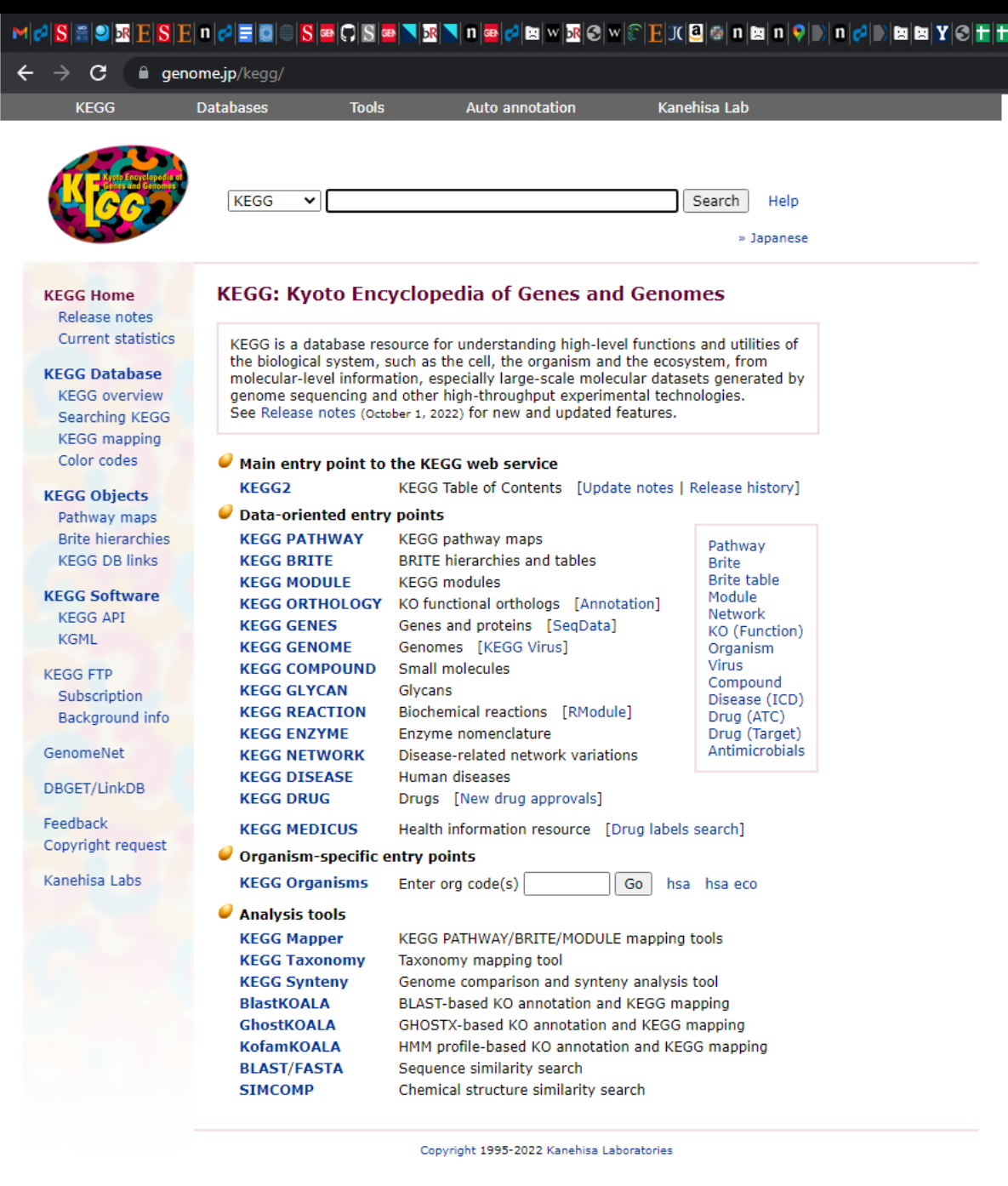
The network of biological classes describing the current best representation of the "universe" of biology: the molecular functions, cellular locations, and processes gene products may carry out.



Statements, based on specific, traceable scientific evidence, asserting that a specific gene product is a real exemplar of a particular GO class.

GO Causal Activity Model (GO-CAM) provides a structured framework to link standard GO annotations into a more complete model of a biological system.

Tools to curate, browse, search, visualize and download both the ontology and annotations. Includes bioinformatic guides (Notebooks) and simple API access to integrate the GO into your research.



- KEGG – Kyoto Encyclopedia of Genes and Genomes
  - http://www.genome.jp/kegg/pathway.html
- a huge database with “pathway” and functional annotation for genes in thousands of genomes

Tryptophan metabolism - Reference pathway - Microsoft Internet Explorer

File Edit View Favorites Tools Help

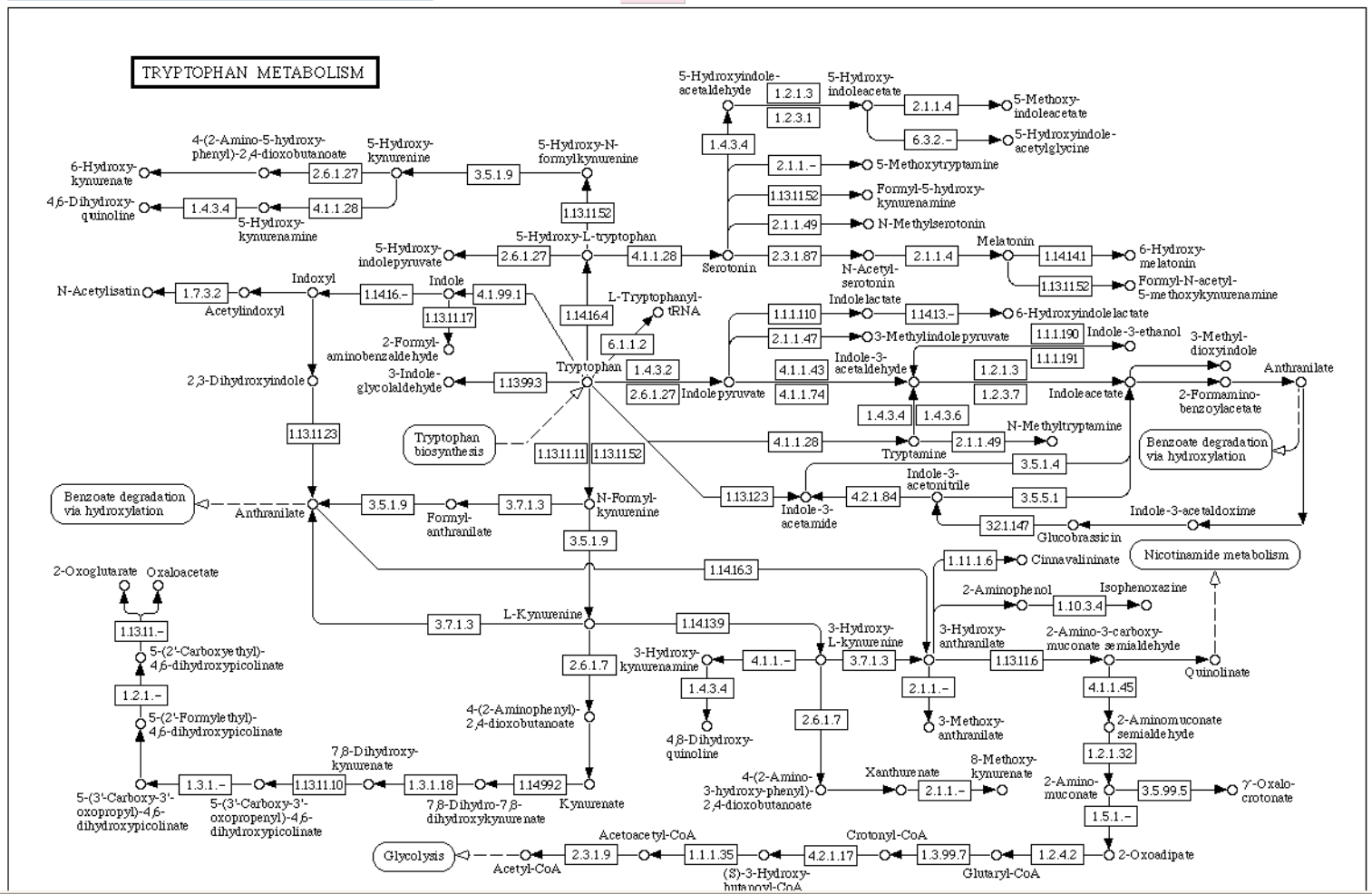
Address http://www.kegg.com/kegg/pathway/map/map00380.html

Tryptophan metabolism - Reference pathway

[Pathway menu]

Reference pathway Go Current selection Select

Help



MSigDB Home

Human Collections
- About
- Browse
- Search
- Investigate
- Gene Families

Mouse Collections
- About
- Browse
- Search
- Investigate

Help

MSigDB
Molecular Signatures Database

Molecular Signatures Database

GSEA and MSigDB Need Your Support

We are preparing a grant proposal to NCI's Information Technology for Cancer Research program for the continued funding of GSEA and MSigDB.

To help ensure their continuing availability, we would greatly appreciate your email of support to accompany our submission. A short email telling us, for example, how you use GSEA/MSigDB in your work, **especially in cancer related projects**, why you chose it over other tools, and your view of its current impact and future promise, will go a long way in underscoring their value of when the grant is reviewed.

Please send your emails to **gsea-los@broadinstitute.org** on or before **Wednesday November 2, 2022.**.

Thanks in advance for your help and support.
The GSEA/MSigDB Team.

UC San Diego

BROAD INSTITUTE

Overview

The Molecular Signatures Database (MSigDB) is a resource of tens of thousands of annotated gene sets for use with GSEA software, divided into Human and Mouse collections. From this web site, you can

- **Examine** a gene set and its annotations. See, for example, the HALLMARK_APOPTOSIS human gene set page.
- **Browse** gene sets by name or collection.
- **Search** for gene sets by keyword.
- **Investigate** gene sets:
  - **Compute overlaps** between your gene set and gene sets in MSigDB.
  - **Categorize** members of a gene set by gene families.
  - **View the expression profile** of a gene set in a provided public expression compendia.
  - Investigate the gene set in the online **biological network repository NDEx**
- **Download** gene sets.

License Terms

GSEA and MSigDB are available for use under these license terms.

Please register to download the GSEA software and the MSigDB gene sets, and

Human Collections

H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** corresponding to human chromosome cytogenetic bands.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **ontology gene sets** consist of genes annotated by the same ontology term.

C6 **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic signature gene sets** represent cell states and perturbations within the immune system.

C8 **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

Mouse Collections

SOME WARNINGS ABOUT ANNOTATION DATABASES

In many cases the definition of a pathway/gene set in a database might differ from that of a scientist

The nodes in pathways are often proteins or metabolites; the activity of the corresponding gene set is not necessarily a good measurement of the activity of the pathway

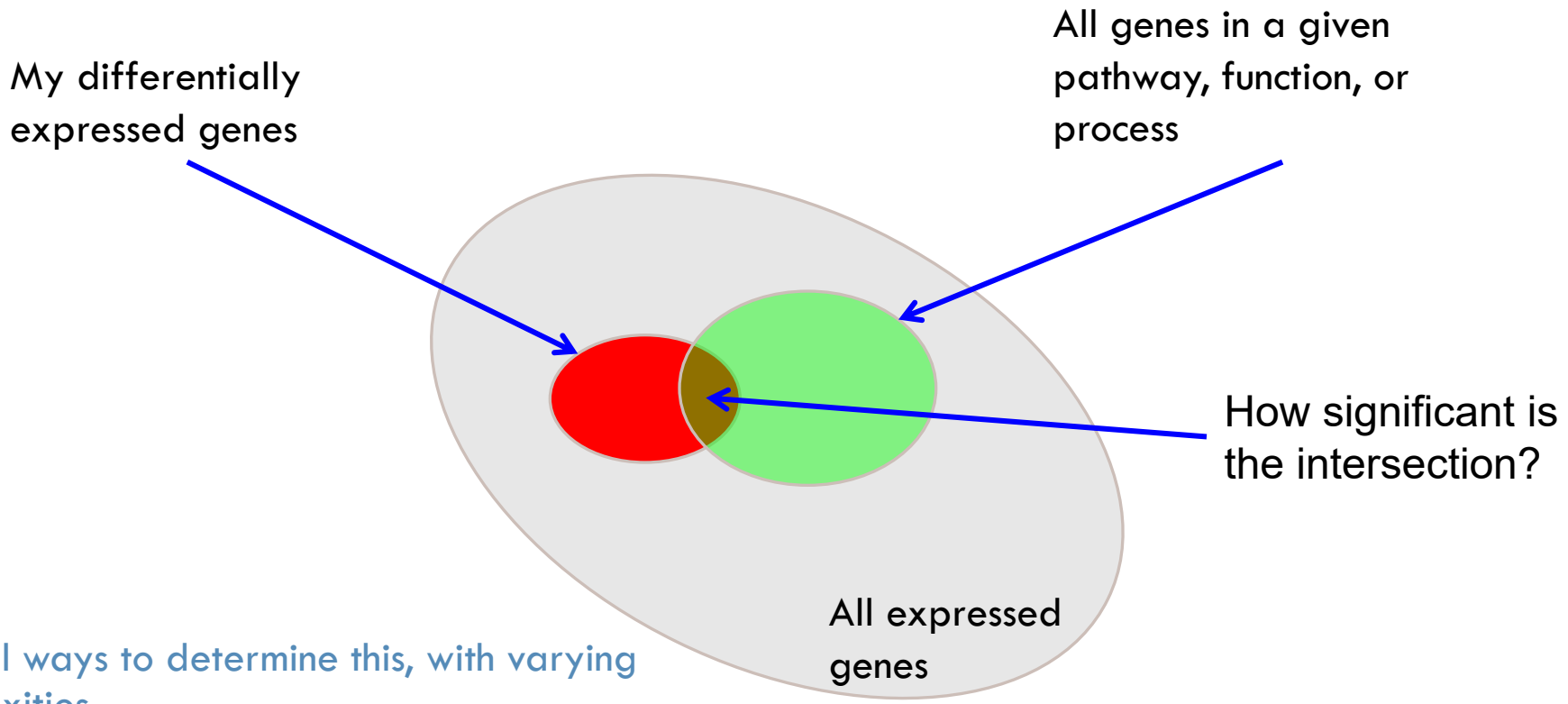
Genes in a gene set often have different identifiers, and sometimes a unique ID maps to multiple genes in another database (Entrez IDs, Unigene IDs, KEGG IDs)

- Conversion can lead to errors!

There are many more resources out there (BioCarta, BioPax, Reactome, Wikipathways, COGs, etc.)

Commercial packages often use their own pathway/gene set definitions (Ingenuity, Metacore, Genomatix,…)

NOW THAT I KNOW WHAT THEY DO, ARE THERE ANY PROCESSES/FUNCTIONS/ETC THAT ARE OVER-REPRESENTED?



Several ways to determine this, with varying complexities

- Hypergeometric test, fisher exact, Chi sq
- GSEA (gene set enrichment analysis)
- Pathway analysis – using network topology

THE HYPERGEOMETRIC DISTRIBUTION

Discrete probability distribution – the prob. Of b successes in B draws from a finite population of size N containing n possible successes

Or, under the probability of finding exactly b black socks in the n randomly chosen socks is described by the hyper-geometric function:

$$HG(N,B,n,b) = \frac{\binom{n}{b}\binom{N-n}{B-b}}{\binom{N}{B}}$$

We are usually interested in the tail probability: finding b **or more** black socks :

$$HGT(N,B,n,b) = \sum_{i=b}^{\min(n,B)} HG(N,B,n,i)$$

THE HYPERGEOMETRIC DISTRIBUTION

- Consider the following scenario:
  - A drawer contains N socks.
  - Exactly B of the socks are black and the remaining (N – B) are white.
  - We pick n socks by random and b of them are black.
- Do the n socks we picked contain significantly more black socks than we expected?
- In other words, are the black socks enriched in the n socks we randomly chose?

THE HYPERGEOMETRIC DISTRIBUTION

Discrete probability distribution – the prob. of only b=1 success in B draws from a finite population of size N containing n possible successes

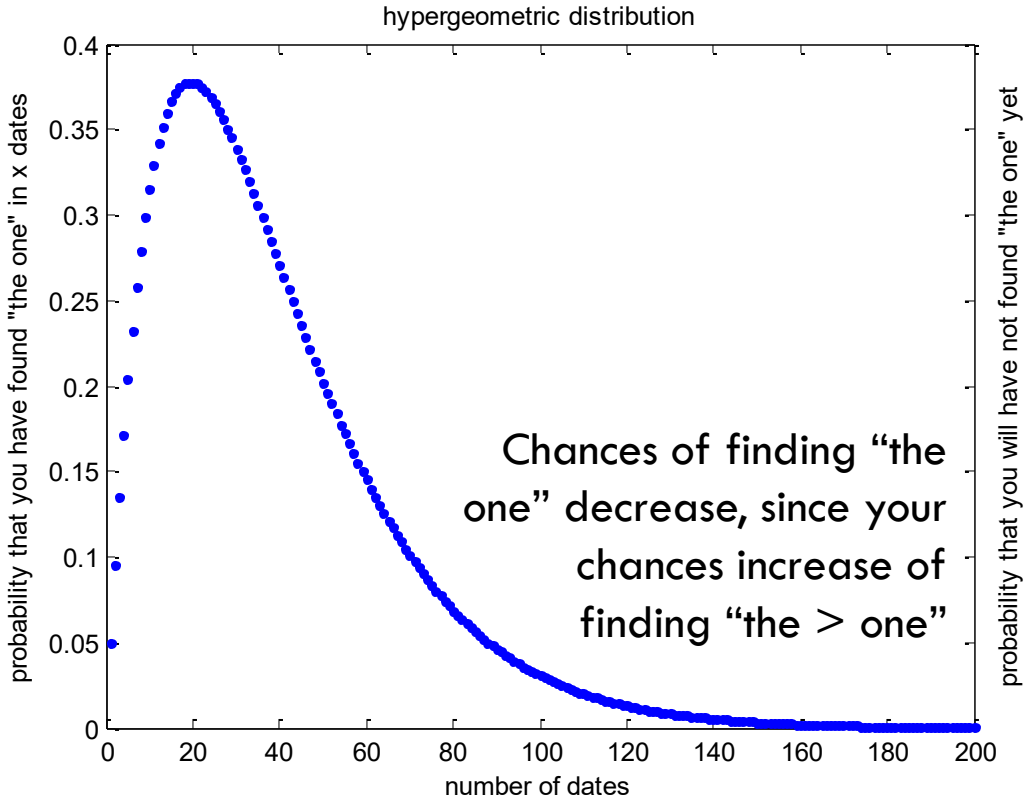
- B = number of dates
- N = Number of singles with matching preference
- n = number that are marriage material
- b = number of marriages… so hopefully 1

$$HG(N,B,n,b) = \frac{\binom{n}{b}\binom{N-n}{B-b}}{\binom{N}{B}}$$

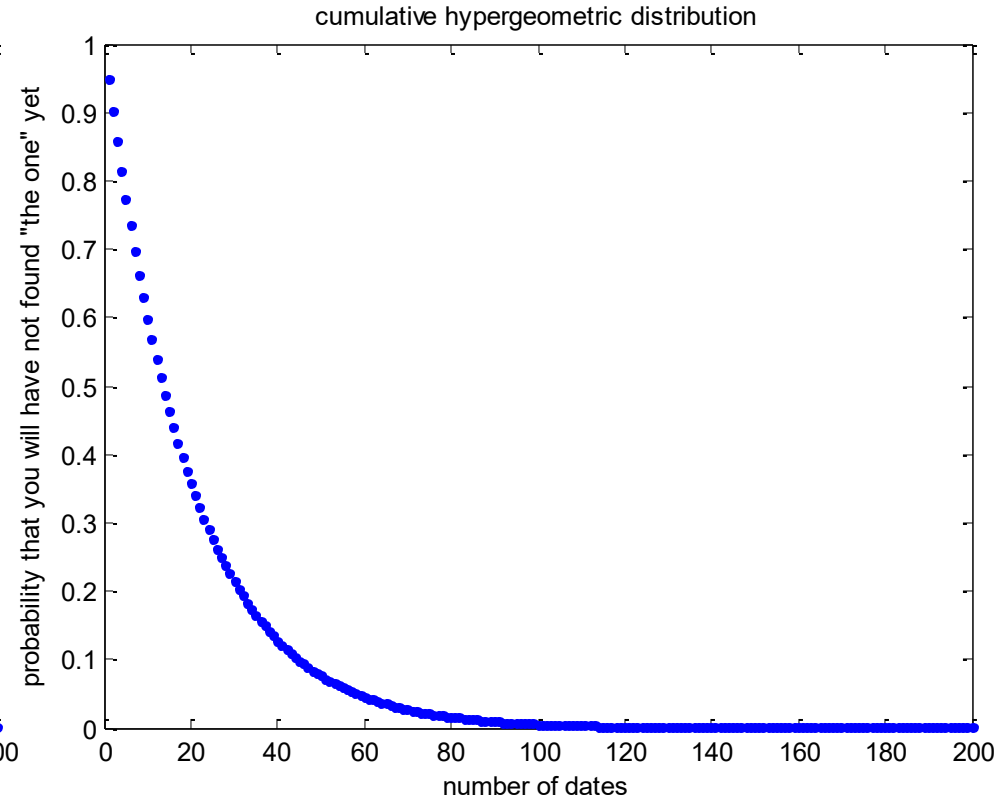
- So, what is the probability of 1 success (i.e., marriage)? (assuming 10000 single people, 5% are marriable, and you ask out any random person, but only once, of course)

HOW MANY DATES WILL YOU NEED?

Not too many…



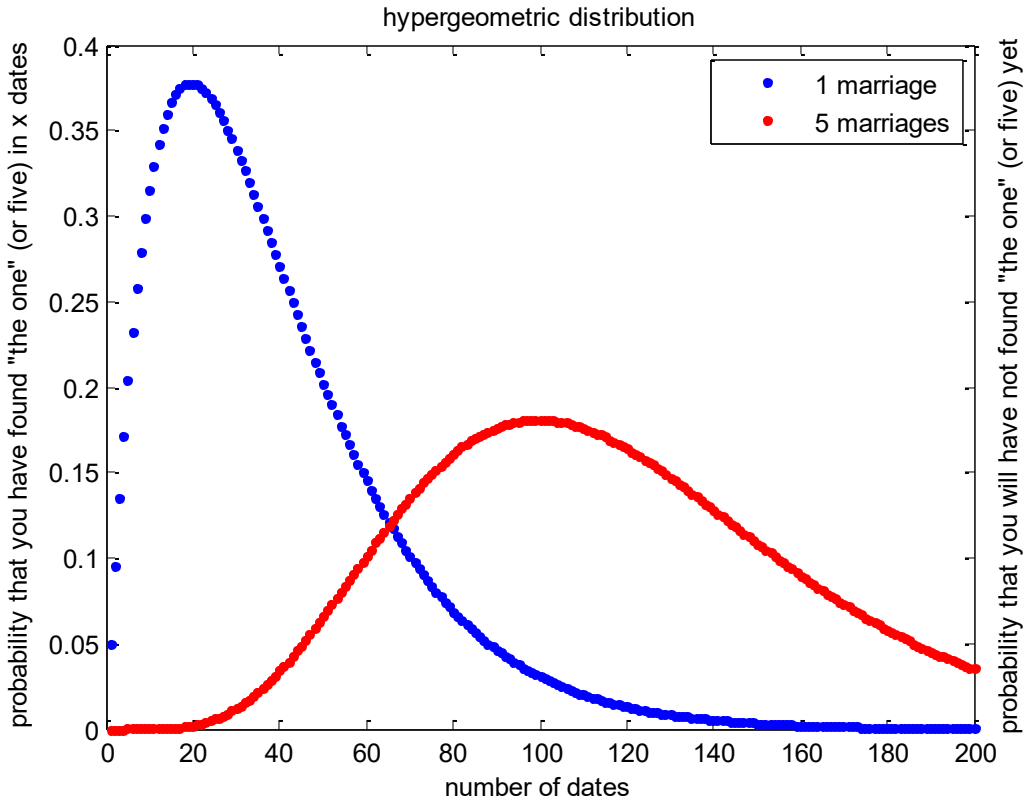
(y is the probability of only 1 success after x dates)

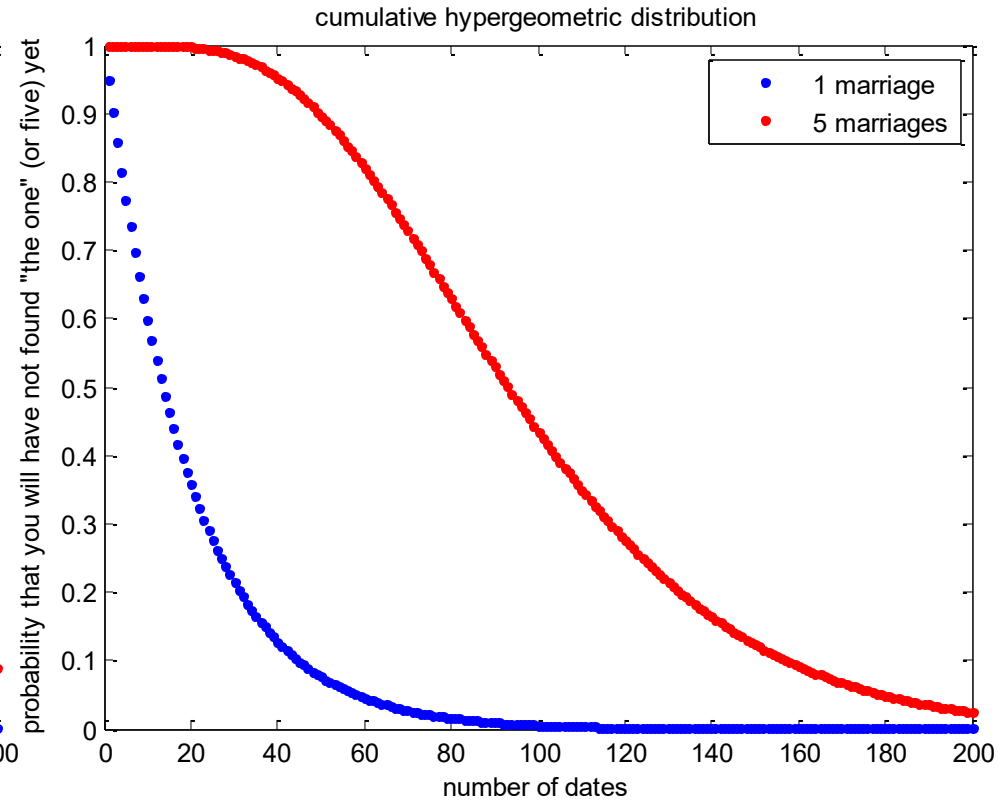


(y the probability of no successes after x dates)

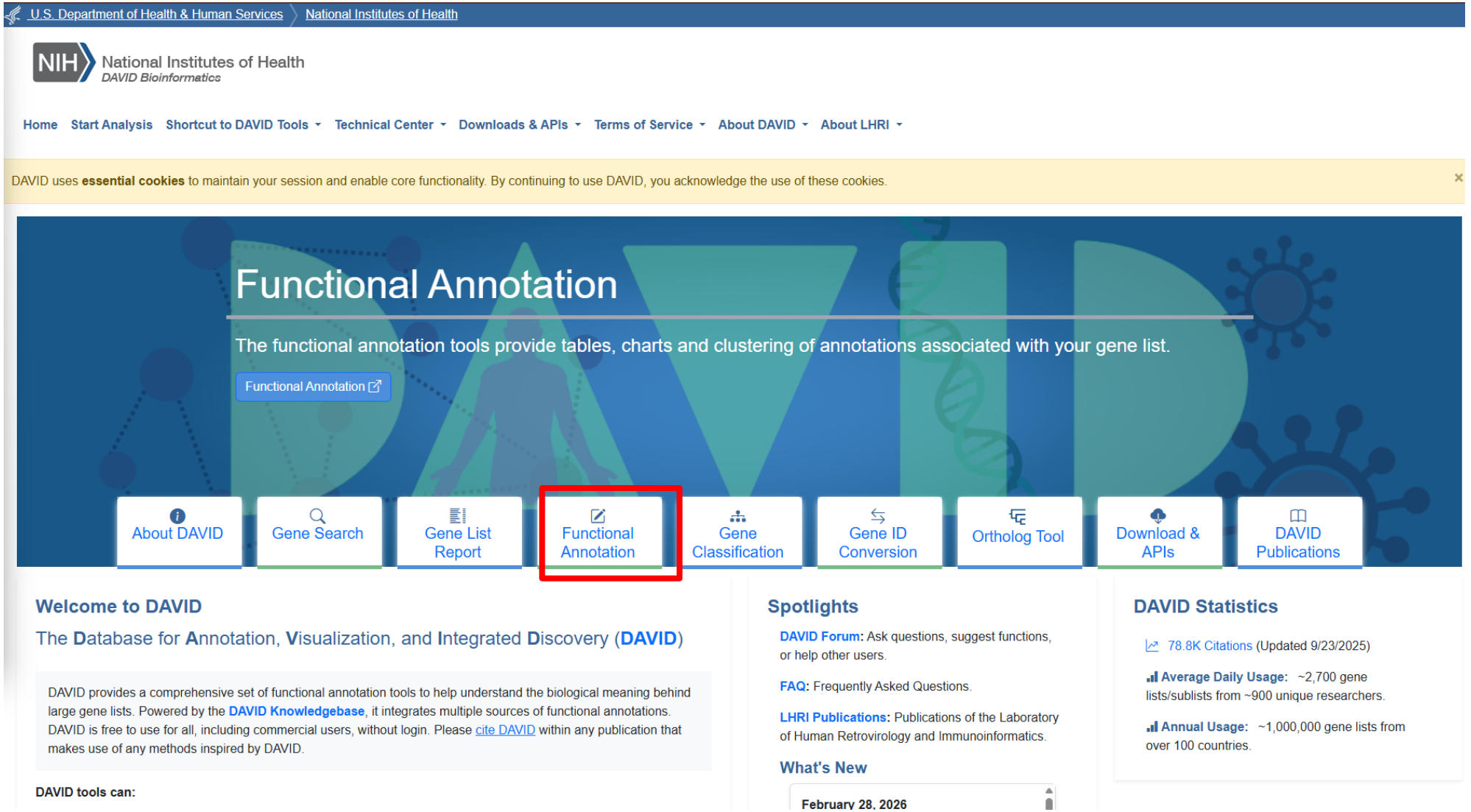
BEST TO STOP AFTER FINDING "THE ONE"… STATS WARN AGAINST TOO MANY DATES

The probability of getting multiple marriages increases





HTTPS://DAVIDBIOINFORMATICS.NIH.GOV/



List

BG

Annotation Summary Results

"significant" genes (all, up, or down)

Current Gene List: MT-CO1
Current Population: Homo sapiens

Clear All | Check Defaults | Functional Annotation Table | Functional Annotation Chart | Functional Annotation Clustering

Disease	2
Functional_Annotations	6
Gene_Ontology	3
General_Annotations	0
Interactions	1
Literature	0
Pathways	1
Protein_Domains	4
Tissue_Expression	0
Transcription_Factors	0

☐ DISGENET	1,818 genes	52.45%	Chart Cluster
☐ GAD DISEASE	2,339 genes	67.48%	Chart Cluster
☐ GAD DISEASE CLASS	2,341 genes	67.54%	Chart Cluster
☑ OMIM DISEASE	927 genes	26.75%	Chart Cluster
☑ UP KW DISEASE	944 genes	27.24%	Chart Cluster

List

BG

Annotation Summary Results

Current Gene List: MT-CO1

Current Population: Homo sapiens

Switch to Classic Versic

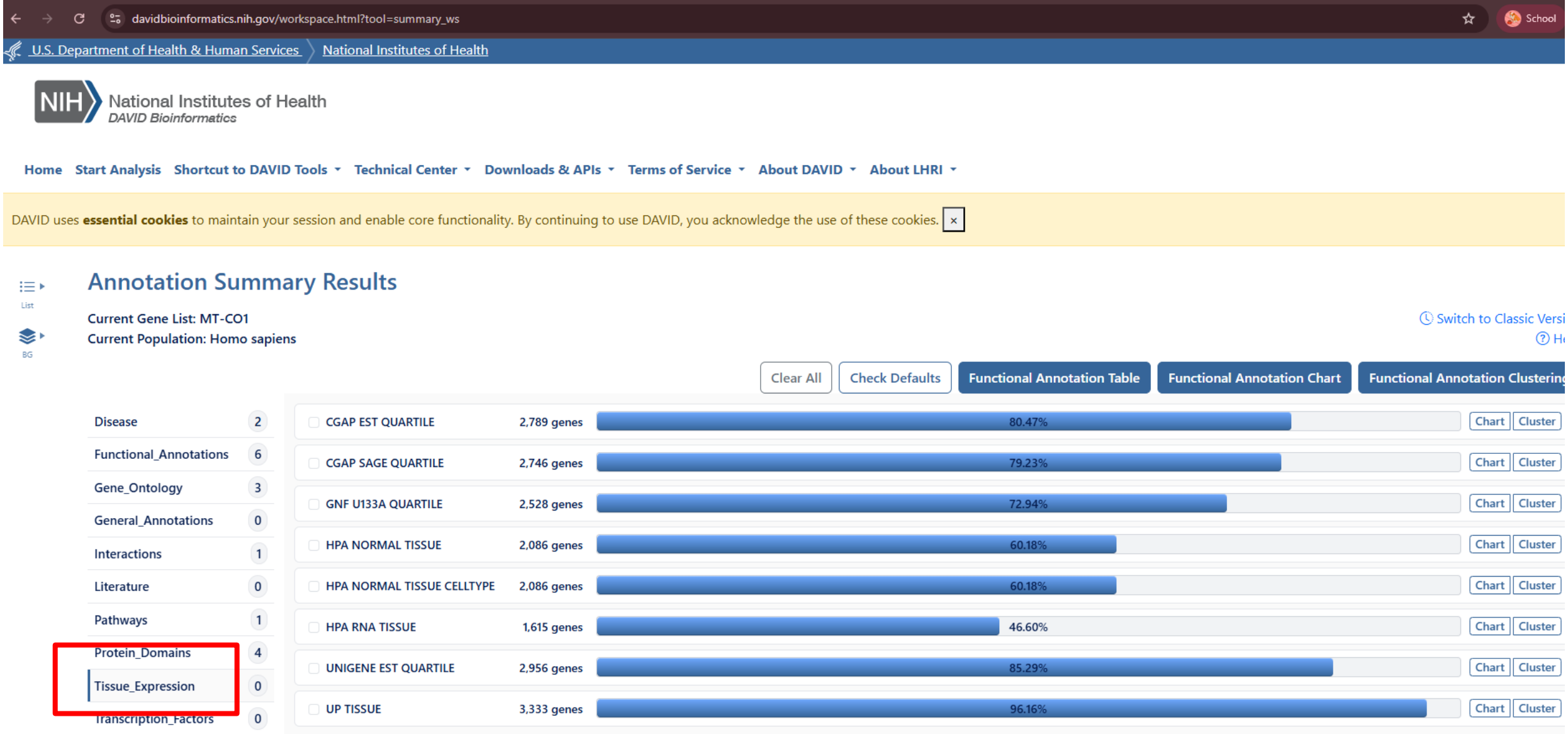
Hel

Background genes (all expressed genes)

Select | DeSelect Defaults | Functional Annotation Table | Functional Annotation Chart | Functional Annotation Clustering

Disease	2
Functional_Annotations	6
Gene_Ontology	3
General_Annotations	0
Interactions	1
Literature	0
Pathways	1
Protein_Domains	4
Tissue_Expression	0
Transcription_Factors	0

DISGENET	1,818 genes	52.45%	Chart Cluster
GAD DISEASE	2,339 genes	67.48%	Chart Cluster
GAD DISEASE CLASS	2,341 genes	67.54%	Chart Cluster
OMIM DISEASE	927 genes	26.75%	Chart Cluster
UP KW DISEASE	944 genes	27.24%	Chart Cluster



Which tissue were they expressed in? (note, only should be using up-regulated genes)

ADVANTAGES OF DAVID

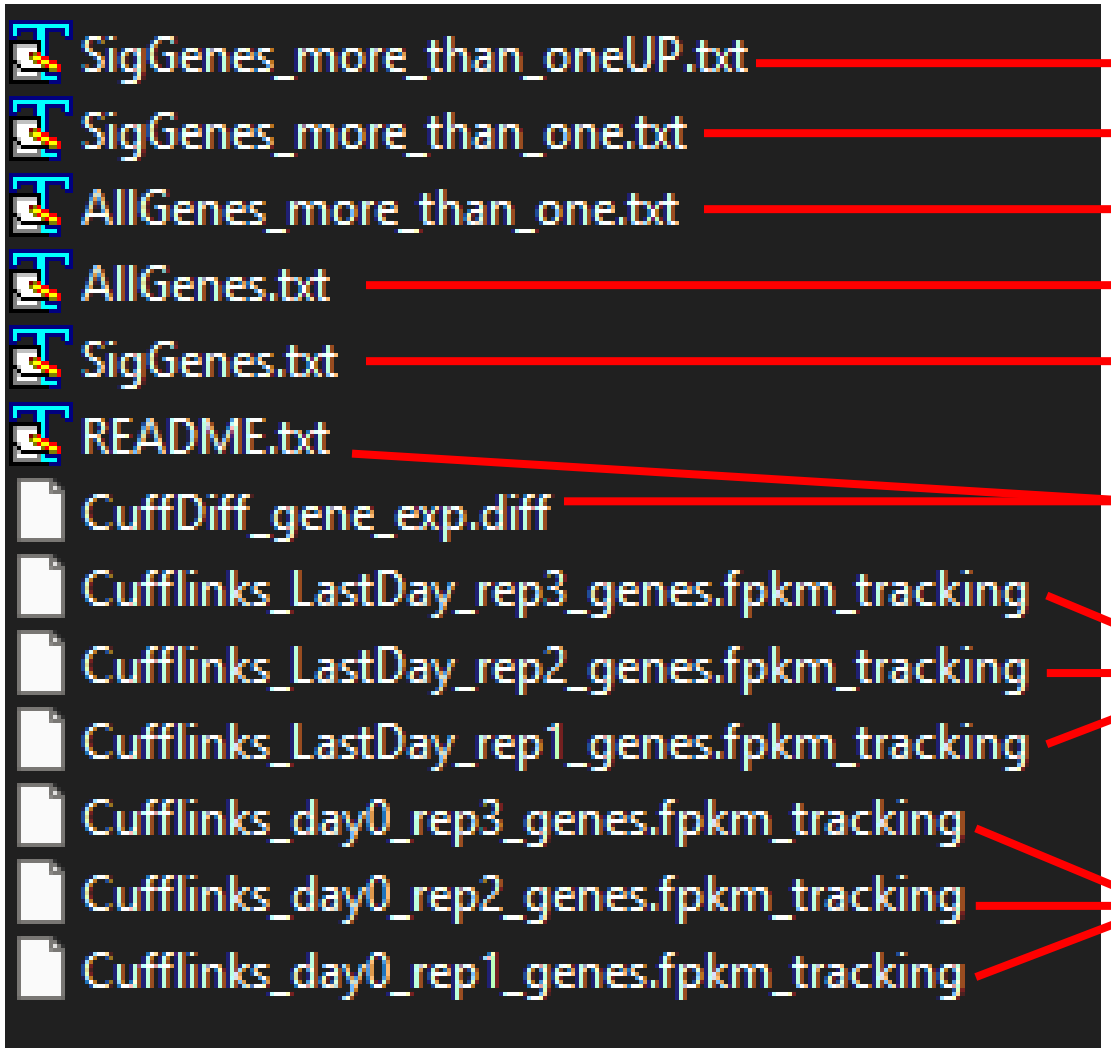
Accept various gene identifiers

Web-based tools automatically retrieve most up-to-date GO annotations

Can often automatically map redundant IDS to a single one, since multiple significant probes for one gene could otherwise skew results

They actually keep updating it… it's not a tool from someone's thesis that is out of date the day it's published ☺

IN CLASS PROJECT



CUT-OFF METHODS VS WHOLE GENE LIST METHODS

A problem with both tests discussed so far is, that they rely on an arbitrary cut-off

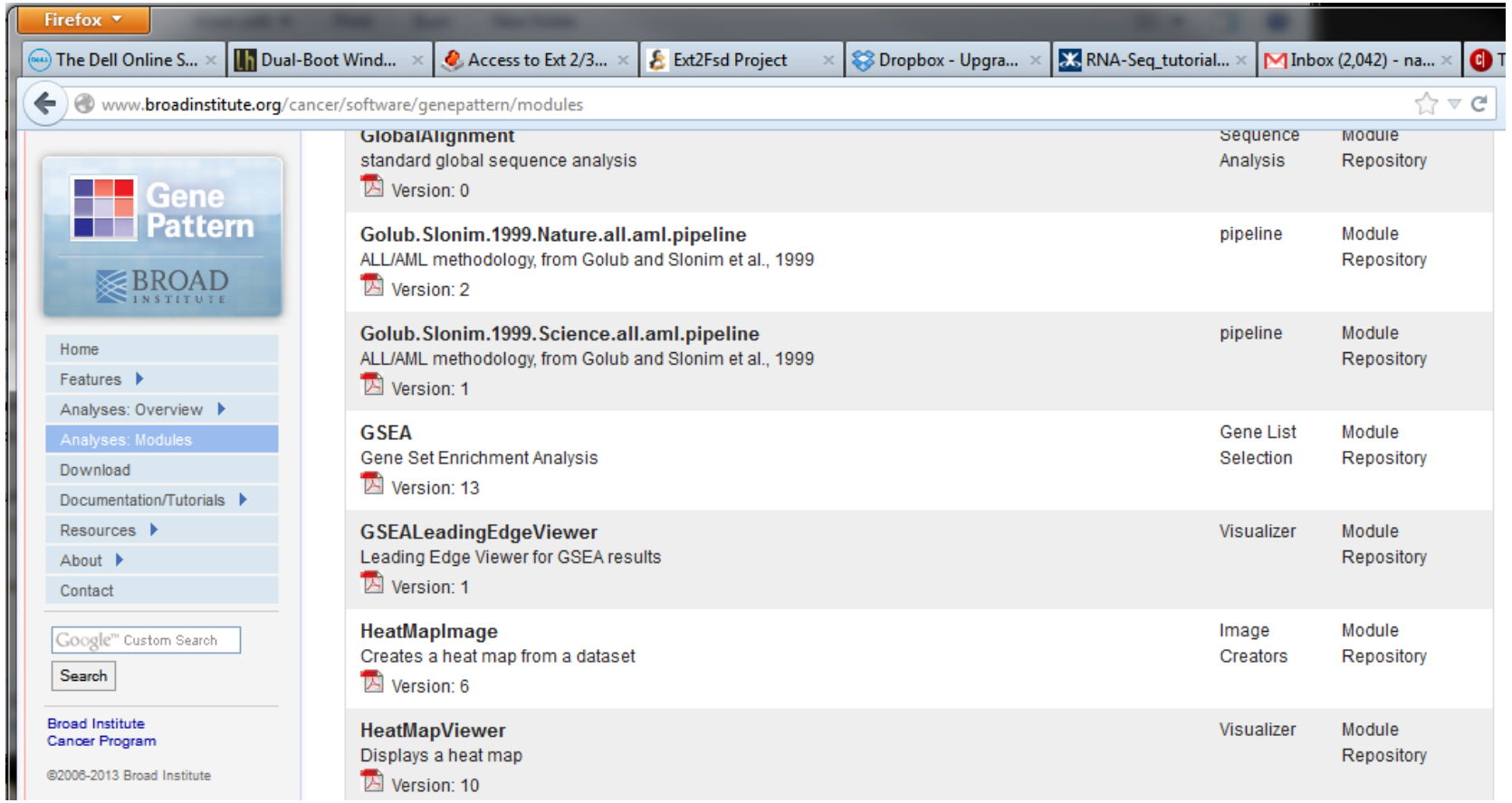
If we call a gene significant for 10% p-value threshold the results will change

- In our example the binomial test yields p= 2.2%, i.e. for this cut-off the result is significant!

We also lose information by reducing a p-value to a binary ("significant", "non-significant") variable

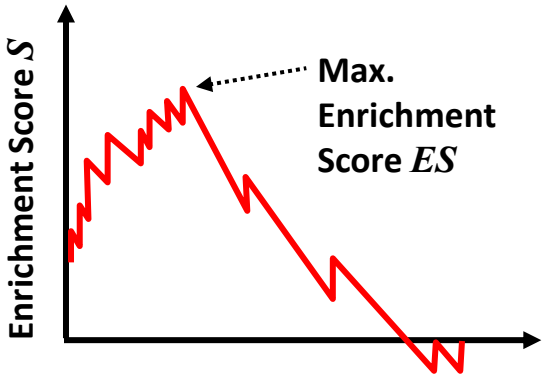
- It should make a difference, whether the non-significant genes in the set are nearly significant or completely unsignificant

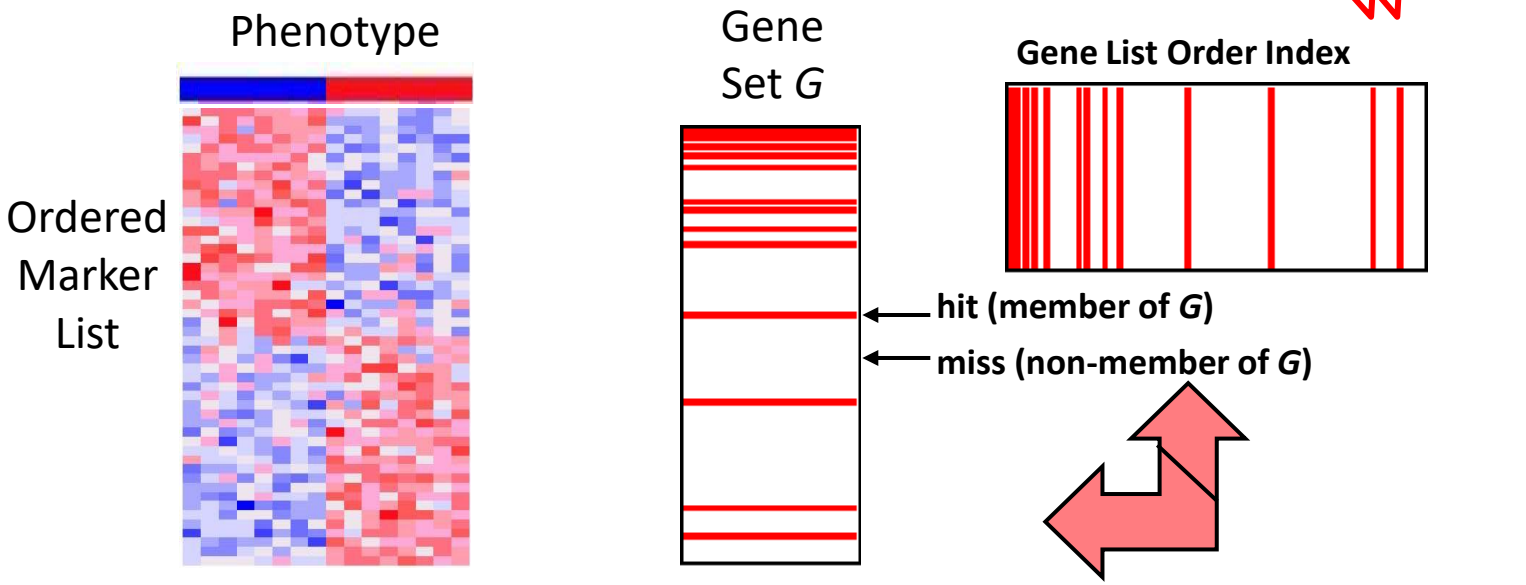
GSEA, A TOOL FOR MORE INTELLIGENT "VARYING" OF THRESHOLDS



ENRICHMENT: KS-SCORE

- Rank genes according to their "correlation" with the class of interest.
- Test if a gene set (e.g., a GO category, a pathway, a different class signature) is enriched.
- Use *Kolmogorov-Smirnoff* score to measure enrichment.

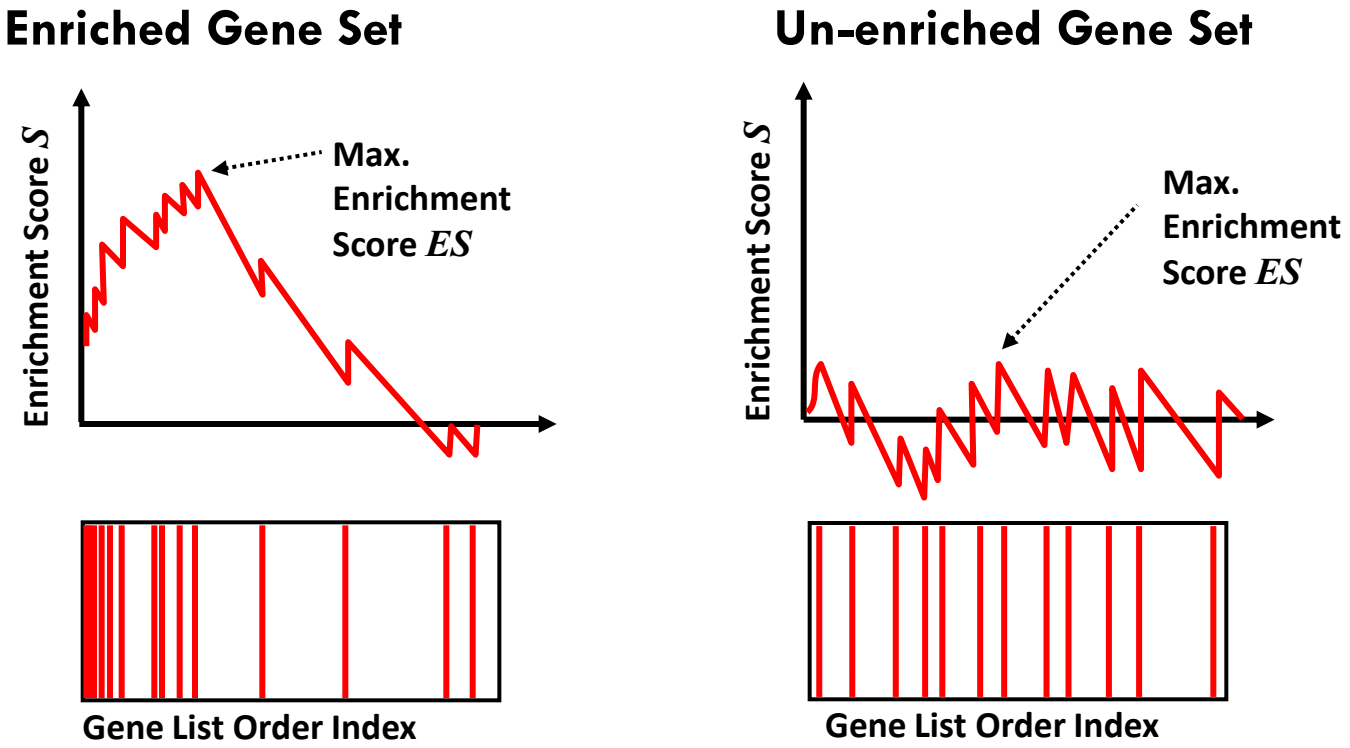




Subramanian et al., PNAS 2005

Mootha et al., *Nature Genetics* 2004

ENRICHMENT: KS-SCORE



Every hit go up by $1/N_H$

Every miss go down by $1/N_M$

The maximum height provides the enrichment score

Running GSEA

1) Use the GenePattern module

2) Use the stand-alone desktop application
   (see www.broadinstitute.org/gsea/downloads)

3) Use the R implementation
   (see www.broadinstitute.org/gsea/downloads)

Inputs

1) Gene expression dataset
   - [or alternatively, a ranked list of genes]

2) Phenotype labels
   - Discrete phenotypes – two or more
   - Continuous phenotypes, e.g. time series

3) Gene sets
   - Select an MSigDB gene set collection
   - Or supply a gene set file

4) Experiment annotations
   - Used to (optionally) collapse expression values into one value per gene
   - Used to annotate genes in the analysis report

Leading edge analysis

- Leading edge subset of a gene set = the genes that appear in the ranked list before the running sum reaches the max value.

- Leading edge analysis = examine the genes that are in the leading edge subsets of the enriched gene sets.
- A gene that is in many of the leading edge subsets is more likely to be of interest than a gene that is only in a few of the leading edge subsets.